Classification Analysis for Environmental Monitoring: Combining Information across Multiple Studies

Huizi Zhang

Dissertation submitted to the faculty of
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirement for the degree of

Doctor of Philosophy
in
Statistics

Eric P. Smith, Chair
Keying Ye
Samantha Bates Prins
Edward Boone

August 14th, 2006
Blacksburg, Virginia

Keywords: Classification, Hierarchical Model, Clustering, Environmental studies

Classification Analysis for Environmental Monitoring: Combining Information across Multiple Studies

Huizi Zhang

Department of Statistics

ABSTRACT

Environmental studies often employ data collected over large spatial regions. Although it is convenient, the conventional single model approach may fail to accurately describe the relationships between variables. Two alternative modeling approaches are available: one applies separate models for different regions; the other applies hierarchical models. The separate modeling approach has two major difficulties: first, we often do not know the underlying clustering structure of the entire data; second, it usually ignores possible dependence among clusters. To deal with the first problem, we propose a model-based clustering method to partition the entire data into subgroups according to the empirical relationships between the response and the predictors. To deal with the second, we propose Bayesian hierarchical models. We illustrate the use of the Bayesian hierarchical model under two situations. First, we apply the hierarchical model based on the empirical clustering structure. Second, we integrate the model-based clustering result to help determine the clustering structure used in the hierarchical model. The nature of the problem is classification since the response is categorical rather than continuous and logistic regression models are used to model the relationship between variables.

Dedication

To my father for his love and inspiration

Acknowledgement

I would like to express my deepest gratitude to my supportive committee members. Especially, I would like to thank **Dr. Eric Smith** for his guidance and support through these years. His openness, insights, and wisdom have been invaluable to my academic development. I am very fortunate to have him as my advisor. Thanks also go to **Dr. Keying Ye** for inspiring me to explore the exciting world of Bayesian Statistics through his excellent teaching. He is also a wonderful friend whom I can always turn to for help and advice. I would like to thank **Dr. Sam Bates Prins** for all the technical help and writing lessons she generously gave me. She is always there when I need her help. Her professionalism and her integrity will have a lasting impact on my future career. I would like to thank **Dr. Ed Boone** for all the thoughtful ideas he shared with me. There were difficult times in my life. It was him who gave me encouragements that kept me moving forward, for which I shall always be grateful.

I would like to thank Dr. Gene Yagow from the Biological Systems Engineering Department at Virginia Tech for helping me drawing all the maps in this dissertation.

I am also deeply indebted to my family members. I am grateful to my mom who helped me take care of my baby during the busiest time of my research. I would like to thank my husband for his love and support without which I could not have made it so far. Above all, I would like to thank my little angel, Michael (葵葵), who has brought so much joy into our life.

Contents

List of Tables

List of Figures

Chapter 1 Introduction

In environmental studies, data containing many variables collected over large spatial regions are often encountered. Statistical models are needed to relate the biological response to some environmental variables. Although it is convenient, the conventional single model approach may fail to accurately describe the relationships between variables. Two alternative modeling approaches are available: one applies separate models for different regions; the other applies hierarchical models. The separate modeling approach has two major problems: first, we often do not know the underlying clustering structure of the entire data; second, it usually ignores possible dependence among clusters. To deal with the first problem, we propose a model-based clustering method to partition the entire data into subgroups according to the empirical relationships between the response and the predictors. To deal with the second, we propose Bayesian hierarchical models. We illustrate the use of the Bayesian hierarchical model under two situations. First, we apply the hierarchical model based on the empirical clustering structure. Second, we integrate the model-based clustering result to help determine the clustering structure used in the hierarchical model. The nature of the problem is classification since the response is categorical rather than continuous and logistic regression models are used to model the relationship between variables.

This dissertation illustrates these methods using a brook trout data collected by the Fish and Aquatic Ecology Unit of the U.S. Forestry Service. The data contains information on the population status of brook trout (*Salvelinus Fontinalis)*, present or extirpated, and various metrics collected in 3333 subwatersheds in the eastern U.S.

This dissertation is organized as follows. In the reminder of this chapter, we will review related topics on logistic regression models, Bayesian hierarchical models, partitioning models and introduce the brook trout data. In Chapter 2, we describe the use of Bayesian hierarchical modeling with an empirical clustering structure. In Chapter 3, we discuss the use of the model-based clustering method in the context of a classification analysis. In Chapter 4, we integrate the model-based clustering results from Chapter 3 to the Bayesian hierarchical modeling approach. Finally, we present extensions and future work in Chapter 5.

1. 1 Logistic Regression

1.1.1 Logistic model for binary response data

Linear logistic regression models belong to the family of Generalized Linear Models (McCullagh and Nelder, 1989). They are widely used in modeling binary data which arise in various fields of study: medical research, pharmaceutical and agricultural experiments, environmental studies, social research, quantitative marketing, to name a few. When one is interested in investigating a binary response (success or failure) for each observation under study, the Bernoulli distribution is usually used as the underlying distribution to do the modeling. If the interest is instead in the proportion of "success" among a group of homogeneous observations, the Binomial distribution is then assumed. The former case is usually referred to as the "ungrouped data" case and the latter the "grouped data" case.

Notationwise, the binary response variable Y is said to have a Bernoulli distribution, denoted as $Y \sim Bernoulli(p)$, if $p(Y = y) = p^y (1-p)^{1-y}$, where p is the probability that Y takes on the value 1 (often referred to as the "success" probability). The variable Y is said to have a Binomial distribution, denoted as $Y \sim Bin(n, p)$, if $p(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$, where n is the total number of independent Bernoulli trials with "success" probability p and y is the number of total successes. It is easily seen that

the Bernoulli distribution is a special case of the Binomial distribution when n equals 1. The general form of logistic model is as follows: Suppose $Y \sim Bin(n, p)$ (or $Y \sim Bernoulli(p)$), the success probability p is modeled using

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \mathbf{x\beta} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

where x_1 through x_m are the explanatory variables, $\boldsymbol{\beta}$ is a vector of regression parameters where the first component corresponds to an intercept term. It is clear that $p = \frac{e^{\mathbf{x\beta}}}{1+e^{\mathbf{x\beta}}} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_m x_m}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_m x_m}}$, hence, the estimated probability under this model is guaranteed to fall into the interval of (0, 1). Another popular model used to model binary data is called the probit regression model. For more details, see Agresti (2002) or Collett (1991).

1.1.2 Estimation method

In classical statistics, the maximum likelihood (ML) method is used to estimate the $\boldsymbol{\beta}$ vector. In Bayesian statistics, inference about $\boldsymbol{\beta}$ is based on the posterior distribution of $\boldsymbol{\beta}$, and is skipped here since we will give a detailed introduction on Bayesian statistics later in the chapter. The likelihood function of a parameter given a sample is the joint probability of the sample under the assumed probability distribution and is denoted as $L(\boldsymbol{\beta})$. It is treated as a function of the parameters and the values of those parameters which maximize this probability are called the Maximum likelihood estimators (MLE). It is usually more convenient to maximize the logarithm of the likelihood, and the MLE of $\log L(\boldsymbol{\beta})$ is the same as the MLE of $L(\boldsymbol{\beta})$. Suppose $\boldsymbol{\beta}$ is of dimension k, the usual scheme is to write k equations of the form:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial(\beta_i)} = 0,$$

where $i = 1,2,...k$, and solve these equations to obtain the MLE of $\boldsymbol{\beta}$. Since there are k non-linear equations, $\hat{\boldsymbol{\beta}}$ can only be solved numerically. Iterative methods such as the

Newton-Raphson method or Fisher Scoring are commonly used. For more computational detail, see Collett (1991). Non-convergence is sometimes encountered due to the iterative nature of the estimating process. Possible causes include complete separation of data or if the observed proportion is always zero or one. When this happens, all estimation results are invalid and should not be used. Several tests are available for testing the significance of the parameters. They are the likelihood ratio test (LRT), the Wald test and the Score test. For more details, see Hosmer and Lemeshow (2000).

1.1.3 Odds ratio

After a logistic regression model is fit, the importance of the parameter β_i is usually interpreted through the odds ratio. The odds of success is defined as $\frac{p}{1-p}$ where p is the success probability. The relative measure of the odds of success under two settings i and j is called the odds ratio and is defined as $\frac{p_i/(1-p_i)}{p_j/(1-p_j)}$. In the logistic model, $e^{\hat{\beta}_i}$ is the odds ratio at x_i+1 versus at x_i, while keeping all other x's constant. In other words, the multiplicative increase in odds when x_i increases by 1 unit is $e^{\hat{\beta}_i}$. In practice, the change in odds ratio for some amount other than one unit may be of greater interest. In that case, people use customized odds ratio. For a change of c units in X_j, the customized odds ratio is estimated by $e^{c\hat{\beta}_i}$.

1.1.4 Model assessment

1.1.4.1 Grouped data case

After a logistic model has been fit to the data, it is important to check the agreement between the fitted values and the observed data. It is referred to as the goodness-of-fit (GOF) test. If the model fits the data well, the model may be acceptable; otherwise, the current model must be revised before use.

In the grouped data case, one tool used in the GOF test is called deviance

$$D = -2[\log \hat{L}_c - \log \hat{L}_f],$$

where $\hat{L}_f$ is then likelihood under the full model and $\hat{L}_c$ is the likelihood under the current model. Suppose we have n binomial observations and $Y_i \sim (n_i, p_i)$, $i = 1,..n$, the deviance is given by:

$$D = 2\sum_i \{y_i \log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i})\},$$

where $\hat{y}_i = n_i \hat{p}_i$. It is known that the deviance is asymptotically χ^2 with $n - p$ degree of freedom, where p is the number of unknown parameters in the model. A small value of deviance usually indicates good fit.

Another popular measure is called the Pearson's χ^2 statistic and it is given as

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}.$$

It is also asymptotically distributed as chi-square with $n - p$ degree of freedom. The two statistics usually give different but similar results. When the difference is large, it may indicate that the chi-square distribution is not a good approximation to the distribution of these two measures.

Deviance is sometimes preferred over the Pearson's χ^2 statistic. One reason is that deviance can also be used to compare two nested logistic models. When comparing two models for binomial data, no exact distribution is available. However the difference between the deviances of two models (denoted as D_1 and D_2 respectively) is also approximately χ^2 distributed. I.e. to compare nested models C_1 and C_2, we use

$$D_1 = -2[\log \hat{L}_{c_1} - \log \hat{L}_f] \dot{\sim} \chi_{n-p_1}{}^2,$$

$$D_2 = -2[\log \hat{L}_{c_2} - \log \hat{L}_f] \dot{\sim} \chi_{n-p_2}{}^2, \text{ then}$$

$$D_1 - D_2 \dot{\sim} \chi_{p_1-p_2}{}^2,$$

where p_1 is the number of unknown parameters in C_1, p_2 is the number of unknown parameters in C_2 and $p_1 > p_2$. A large value of this statistic indicates the more complex model (C_2 in above case) is preferred over the simpler one.

1.1.4.2 Ungrouped data case

For ungrouped data, it is shown that (Collett, 1991)

$$D = -2\sum_i \{\hat{p}_i \text{logit}(\hat{p}_i) + \log(1-\hat{p}_i)\}.$$

Clearly the deviance depends only on the predicted probability $\hat{p}$. It says nothing about the agreement between fitted values and data; therefore it cannot be used to test goodness-of-fit. An alternative test called the Hosmer-Lemeshow test was developed. The basic idea is to divide data into *g* approximately equal size groups based on the percentiles of the estimated probabilities. The discrepancies between observed and fitted values are summarized by the Pearson Chi-square statistics as the following:

$$\mathrm{X}_{hl}^{\ 2} = \sum_1^g \frac{(o_i - n_i\overline{\pi}_i)^2}{n_i\overline{\pi}_i(1-\overline{\pi}_i)}$$

where n_i is the number of observations in the *i*th group, o_i is the number of event outcomes in the *i*th group and $\overline{\pi}_i$ is the average estimated probability of an event outcome for the *i*th group. This statistic is compared to an asymptotic $\chi_{g-2}^{\ 2}$ distribution to test for GOF. A common choice of *g* is 10. For this test to work well, a moderate cell count is required after grouping. For *g*=10, the rule of thumb is that the cell count be at least 5. For more details about this test, see Hosmer and Lemeshow (2000). It has been shown that this test is conservative. It has low power of detecting specific types of lack of fit such as nonlinearity in the explanatory variable and it is highly dependent on how the observations are grouped (SAS Institute Inc, 1995).

1.1.5 Predictive accuracy

1.1.5.1 AFCCF

Often in classification studies, one is more interested in the predictive accuracy of a model. For ungrouped data, two measures are available for assessing the predictive power of a model. The first one is called Average Fraction of Correctly Classified for Fit (AFCCF) (Birch, 2002, SYSTAT® 9). It is defined as

$$\text{AFCCF}=\frac{\sum_{1}^{n} y_i \hat{y}_i + \sum_{1}^{n}(1-y_i)(1-\hat{y}_i)}{n},$$

where n is the sample size, y_i is the binary response, $\hat{y}_i$ is the fitted value for the response ($0 \le \hat{y}_i \le 1$). A larger value indicates better predictive power of the model.

1.1.5.2 Sensitivity, specificity and AUC (Area under the ROC curve)

A natural question to ask in logistic regression is whether an observation is classified as 1 (success, presence), or 0 (failure, absence). Since the fitted value for an observation $\hat{y}_i$ is a number between 0 and 1, we need a cutoff value to decide whether to classify this observation as 1 or 0. A natural and most commonly used cutoff is c=0.5. When $\hat{y}_i \ge 0.5$, $\hat{y}_i$ is classified as 1, otherwise, it is classified as 0. To assess a model's classification ability, three measures are commonly computed. The overall correct classification rate is the percentage of observations correctly classified out of all the data points. Sensitivity is the percentage of observations correctly classified as 1 among all responses of 1. Specificity is the percentage of observations correctly classified as 0 among all responses of 0. See Table 1.1 for the illustration. It is clear that these measurements depend on the cutoff value c and there is a tradeoff between sensitivity and specificity.

Another good way to evaluate a model's discriminating power is through the use of the ROC curve. The Receiver Operating Characteristics (ROC) curve was developed in the 1950s in radio signal detection. The ROC curve is a graphic display that gives a measure

of the predictive (discriminant) accuracy of a classification model. The area under the ROC curve (AUC) is the most common summary index describing an ROC curve and has long been used as a measure of classification performance (Bamber, 1975, Hanley and McNeil, 1982, Hanley, 1998, Pepe, 2000, 2003, 2005, Ma and Huang, 2005). The usual estimator for this area can be written as

$$\mathrm{AUC} = \frac{1}{n_S n_F} \sum_{i \in S, j \in F} I(p_i > p_j)$$

where S and F are the index sets for success and failure groups with size n_S and n_F respectively; p_i and p_j are the predicted probability of success for the ith (jth) observation in the success (failure) group and I corresponds to the indicator function (the indicator function assigns the value one if the condition is true). It is interesting to note that AUC is closely related to the Mann-Whitney statistic for two sample problems, which is equivalent to the Wilcoxon rank statistic (Bamber, 1975). A larger value of AUC indicates stronger discriminating power of the model. An area of 0.5 indicates that the model has no predictive power and 1 indicates perfect prediction.

Table 1.1: Illustration of sensitivity and specificity

		Fitted value	
		1	0
True	1	A	B
Response	0	C	D

Sensitivity=A/(A+B)

Specificity=D/(C+D)

Overall correct rate=(A+D)/(A+B+C+D)

1.1.6 Infinite parameters

The likelihood equation for a logistic regression mentioned earlier does not always have a finite solution. The term "infinite parameters" are used to refer to such situations when no unique maximum likelihood estimates for the parameters exist (Albert and Anderson, 1984). In other words, the estimation process fails to converge. This occurs when the data display certain pattern, such as "complete separation" or "quasicomplete" separation (So, 1993). Some possible remedies include examining original data for errors, use fewer or different explanatory variables, rescale the explanatory variable and collect more data.

1.2 Model-based clustering

1.2.1 Introduction

Researchers sometimes are faced with data collected from different sources or over different regions. Using a single parametric model in such data analysis may not be appropriate to describe the relationship between variables. It is still common practice since it is easy to implement. However, increased computing power has enabled researchers to develop data-adaptive methods that are more flexible than the conventional single model approach. Some examples include Multivariate Adaptive Regression Splines (Friendman, 1991), Classification and regression trees (Breiman, et al., 1984), Local polynomial models (Fan and Gijbels, 1996) and Neural networks (Bishop, 1995). Two more recent methods are the Bayesian Partitioning Model (Holmes, et al., 1999) and the Bayesian Treed Model (Chipman, 2002). Bayesian Partitioning Models (BPM) is based on the idea of dividing the design space into a set of homogeneous regions. Within each region, they assume the data come from the same normal or multinomial distribution. Conjugate priors are used for these models so that the marginal distributions of these models can be obtained analytically. The number of regions and the boundaries are assumed unknown *a priori* and Markov chain Monte Carlo techniques are used to obtain the posterior structure of the partition. The final inference is based on the model averaging result of all posterior models. They further extend this method to deal with analysis of spatial count data (Denison and Holmes, 2001). Bayesian Treed Model (BTM)

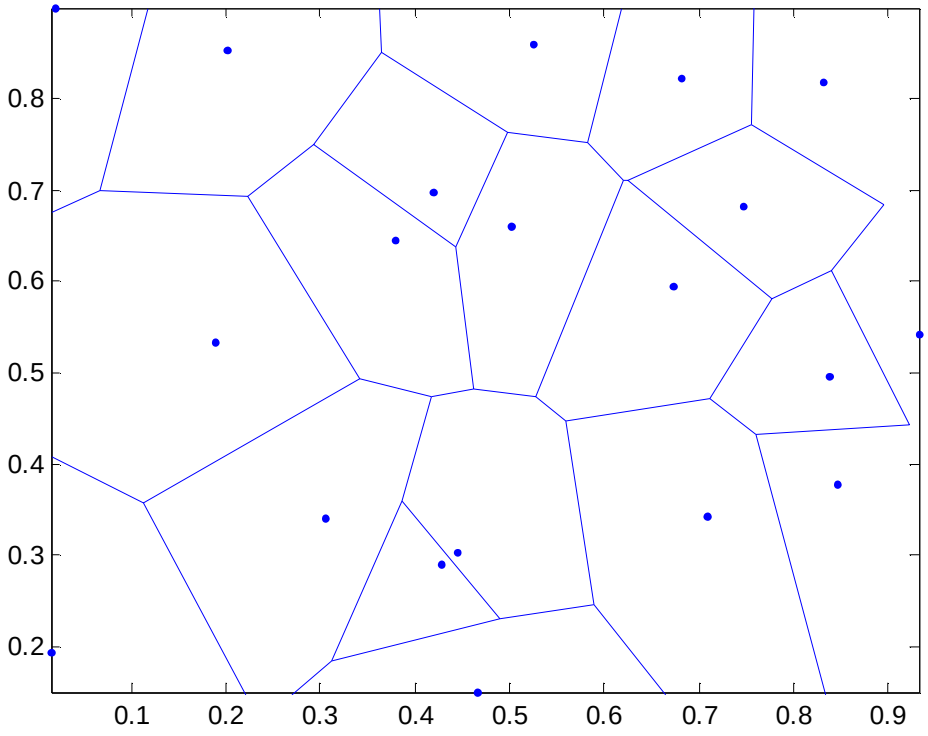
is an extension of the conventional trees model (e.g. CART) that uses binary trees to partition the data into homogeneous subset in a Bayesian framework. It is constructed with the aim of homogeneous "model structure" within each terminal node. Like BPM, it assumes a prior distribution on both the tree structure and the models. Because of the way the partition is done, this method allows only "axis-parallel" partition, producing clusters of rectangular shape. The same variables are used in both partitioning the design space and fitting the terminal node model.

1.2.2 Voronoi tessellations

A special technique used in partitioning the whole space into subregions is called Voronoi Tessellations. The history of Voronoi tessellations goes back to Dirichlet (1850) who used a special form of the Voronoi tessellation in a study of positive definite quadratic forms. Later Voronoi (1908), a Russian mathematician, generalized this to higher dimensions. Since then, Voronoi tessellations have been discovered in diverse scientific disciplines, different names are used such as Dirichlet regions, Thiessen polygons, Voronoi diagrams, Voronoi tessellation, etc (Møller, 1994). This is a rapidly expanding topic as these techniques find applications in such areas as spatial data manipulation, modeling spatial structures and spatial processes, pattern analysis and locational optimization (Okabe, et al., 2000).

A Voronoi tessellation can be formed in the following way. Consider n finite points (labeled $p_1,...,p_n$) and a fixed discrete set of k reference points *(labeled* $S_1,...,S_k$) in Euclidean space. For almost any point p, there will be a reference point that p is closer to than to any of the other reference points based on certain distance measure. The Euclidean space is then decomposed by assigning each point to one of the k reference points to which it is closest, forming a Voronoi tessellation. See Figure 1 for an illustration in a two-dimensional space.

Figure 1.1: Illustration of a two-dimension Voronoi tessellation with 20 reference points



Some earlier mathematical and statistical applications of Voronoi tessellations include packing theory (Rogers, 1964), spatial interpolation of surfaces (Sibson, 1980) and analysis of spatial point patterns (Ripley, 1981). Møller (1994) gives rigorous mathematical details and proofs on the use of random Voronoi tessellations in statistics. A few recent application using Voronoi Tessellations in statistical problems include a disease risk study (Denison and Holmes, 2001), ecological regression studies of geographical epidemiology (Greco, et al., 2005) and soil permeability studies (Kim, et al., 2005). The methods we propose later also utilize these techniques.

1.2.3 Model-based clustering

1.2.3.1 Introduction

The Model-Based Clustering method we propose is based on the idea of partitioning models mentioned earlier in the introduction. The general method we present here is based on the approach of Prins et al. (unpublished document). Several features of the method include 1). It clusters data by the relationship between the explanatory variables and the response variable. 2). The clustering variables are different from the explanatory variables and are usually natural covariates contained in the data. 3). A randomization approach is used in decomposing the entire data into subgroups. 4). It is flexible in terms of the types of data it can handle (e.g. continuous, categorical and multivariate). 5). It is flexible in terms of the "optimality" criteria one can use to reflect various research goals.

1.2.3.2 Implementation

The actual steps to implement this method are the following:

Step1. For a given *k*, form *k* random clusters. We generate k random seed points and assign each data points to one of the *k* random seed points forming *k* random clusters. In environmental studies, data are collected spatially forming a continuous space. Therefore we use the Voronoi tessellations to form cluster boundaries for the entire region.

Step 2. Fit a regression model in each cluster. Given a clustering of the data points, we use a parametric model to fit the clustered data. For continuous response data, linear models are usually used. For categorical (binary) data, logistic models are used. For multivariate data, multivariate regression models are used. In fact, one can assume other types of regression models depending on the types of studies.

Step 3. Evaluate the quality of current clustering solution. Depending on the nature and purpose of the study, "optimality' criteria are selected and computed to assess the current clustering solution. Likelihood criteria with some penalties are used on linear models and

classification performance criteria. Specifically, the area under the ROC (AUC) is used in classification analysis.

Usually, we will explore various choices of k (typical choices include 2 through 10), and we repeat the previous steps a large number of times. The final decision in selecting the clustering solution to use is a combination of the numerical performance in terms of the optimality criteria combined with subject knowledge.

1.2.3.3 Sensitivity analysis

Since our method uses a Monte Carlo search process to locate the "best" clustering solution, naturally, we would need to vary the starting points for this simulation process and run the simulations for a large number of times to ensure that we have achieved some degree of "convergence" in terms of the optimality measure each chain produces and the actually clusters they form. We do admit that as the number of observations increase, it is computationally infeasible to do an exhaustive search on all the possible clustering schemes.

1.3 Bayesian hierarchical model

1.3.1 Bayesian vs. Frequentist

There are two major philosophically different approaches to statistics: the Bayesian approach and the Frequentist approach. One essential difference between them is their view toward an unknown parameter of interest: θ. (Here θ is used as a general notation to denote scalar, vector or matrix depending on the context of problem). While the Frequentist treats θ as fixed and their methods are based on hypothetical infinite repetitions of experimental results conditional on the unknown θ. The Bayesian, instead, views θ as a random variable and tries to understand θ through a probability distribution conditional on the observed data (Berger, 1985). The Bayesian allows the use of prior information whenever available to help with posterior inference about θ while the

Frequentist discards such prior information to maintain so called "objectiveness" (Berger, 1985). The Bayesian's common-sense interpretation of probability as a direct measure of uncertainty about a parameter estimate is one reason that it is favored over the Frequentist by many practitioners (Berger, 1985, Gelman et al., 1995).

1.3.2 Introduction to hierarchical modeling

Data with hierarchical structure are often encountered in scientific applications. Parameters governing each hierarchy level are usually not independent due to the hierarchical structure and therefore a joint probability distribution is assumed for these parameters to reflect their dependence. This is the basic idea of Bayesian hierarchical modeling (Kass and Steffey, 1989).

The general Bayesian hierarchical model can be written as follows (Carlin and Louis, 1996).

$$
\begin{aligned}
Y \mid \theta &\sim f(y \mid \theta), \\
\theta \mid \eta_1 &\sim \pi_1(\theta \mid \eta_1), \\
\eta_1 \mid \eta_2 &\sim \pi_2(\eta_1 \mid \eta_2), \\
&\ldots \\
\eta_{L-1} \mid \eta_L &\sim \pi_{L-1}(\eta_{L-1} \mid \eta_L), \\
\eta_L &\sim h(\eta_L),
\end{aligned}
$$

where Y is the data, and $\theta, \eta_1, \ldots \eta_L$ are the parameters. Usually θ is the parameter of interest, $\eta_1, \ldots \eta_L$ are the hyper-parameters which influence the data Y only through θ. $Y \mid \theta$ is distributed according to $f(y \mid \theta)$, $\theta \mid \eta_1$ is distributed according to $\pi_1(\theta \mid \eta_1)$, and so on up to η_L with the hyperprior distribution to be $h(\eta_L)$.

Questions arise as to how many levels of hierarchy one should explore. The answer depends on factors including one's available resources (time, money) and the degree of satisfaction with the current model. By the law of diminishing returns, models with 2-4 layers should suffice in most cases.

By Bayes' Rule, the posterior distribution of θ is computed as

$$p(\theta \mid y) = \frac{\int ....\int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)...h(\eta_L)d\eta_1 d\eta_2 ...d\eta_L}{\int ....\int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)...h(\eta_L)d\theta d\eta_1 d\eta_2 ...d\eta_L}.$$

Such modeling strategies, either from a fully or empirical Bayesian perspective, helps deal with the multi-parameter problem, usually improves precision of each individual parameter estimate through "pooling strength" and performs better in terms of predictive ability (Gelman et al., 1995, Congdon 2001, Congdon 2003). Since its introduction by Good (1965), it has attracted increased interest from many statisticians. Box and Tiao (1973) study in detail the variance components model in the normal case. Berger (1985) illustrates the use of such models in the normal case and discusses hierarchical priors which Lindley and Smith (1972) refer to as a multistage prior. Laird and Ware (1982) illustrate its use in random-effects models and the modeling of longitudinal data. Meng and Dempster (1987) use this approach in toxicology. Wong and Mason (1985) study the use of modern contraceptive methods in developing countries. Belin et al. (1993) study undercount estimation problem in census data. Gelman and Little (1997) employ this method to study U.S election poll data. Kahn and Raftery (1996), Bedrick, Christensen and Johnson (1997), Daniels and Gatsonis (1999) use the model in medical research areas. Clayton and Kaldor (1987), Richardson and Gilks (1993) find its use in epidemiology. Browne and Draper (2004) and Scott and Ip (2002) apply this method in educational studies. Examples of applications in ecological settings include those of He and Sun (1998, 2000) and Wolpert and Warren-Hicks (1992).

The validity of the hierarchical model depends on the exchangeability assumption. Suppose we have a set of data $y_j \mid \theta_j, j = 1,2,..,J$. When no information other than the data is available, we could not tell which θ_j governs the generation of data y_j, therefore we assume the θ_j 's are symmetric. In other words, parameters $(\theta_1, \theta_2, ...\theta_J)$ are exchangeable if their joint distribution $p(\theta_1, \theta_2 ...\theta_J)$ is invariant to the permutation of the index $(1,2,..,J)$ (Gelman et al., 1995).

$$p(\theta_1, \theta_2 ... \theta_J \mid \phi) = \prod_{i=1}^{J} p(\theta_i \mid \phi) .$$

ϕ is generally unknown and we use integration to obtain the distribution of θ:

$$p(\theta) = \int \prod_{i=1}^{J} p(\theta_i \mid \phi) p(\phi) d(\phi) .$$

Notice this exchangeability assumption is weaker than the independent and identically distributed (*iid*) assumption. IID implies exchangeability but exchangeable θ_j's are not necessarily independent (Browne and Draper, 2004).

1.3.3 Empirical Bayes method vs. fully Bayesian method.

1.3.3.1 Empirical Bayes (EB) method

Recall that the posterior distribution of the parameter θ is written as

$$p(\theta \mid y) = \frac{\int ....\int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)...h(\eta_L) d\eta_1 d\eta_2 ... d\eta_L}{\int ....\int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)...h(\eta_L) d\theta d\eta_1 d\eta_2 ... d\eta_L}$$

In the Empirical Bayes approach, the hyper-parameter η_L in (1.1) is estimated by maximizing the marginal distribution of *Y*:

$$p(Y \mid \eta_L) = \int ....\int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)...\pi_{L-1}(\eta_{L-1} \mid \eta_L) d\theta d\eta_1 d\eta_2 ... d\eta_{L-1} .$$

This estimate $\hat{\eta}_L$ is then plugged into the posterior distribution of θ, and the inference of θ is now based on $p(\theta \mid y, \hat{\eta})$. This simplification greatly reduces computational burden, which, in some cases is vital to problem-solving. One disadvantage of this EB method is its failure to incorporate the uncertainty about this hyper-parameter η_L into the model resulting in an underestimate of the posterior variance (Gelman et al., 1995, Scott and Ip 2002). Kass and Steffey (1989) illustrate it in a two-stage model as follows: Suppose there are k units, the observation vectors, y_i, for units $i = 1,2,...k$, are independently distributed as $p(y_i \mid \theta_i)$ and the unit-specific parameter vector θ_i are independent and identically distributed with density $p(\theta_i \mid \lambda)$. When the EB method is used, inference

about θ_i is based on the posterior distribution $p(\theta_i \mid y_i, \hat{\lambda})$. We write the posterior variance of θ_i as

$$Var(\theta_i \mid y) = E_\lambda\{Var(\theta_i \mid y_i, \lambda)\} + Var_\lambda\{E(\theta_i \mid y_i, \lambda\}.$$

The conditional posterior variance $Var(\theta_i \mid y_i, \hat{\lambda})$ approximates only the first term and therefore is less than $Var(\theta_i \mid y)$. The posterior mean of $\theta_i \mid y, \hat{\lambda}$ is a first order approximation to the posterior mean of $\theta_i \mid y$ and generally speaking, the approximation is often good. The above mentioned approach is also referred to as Parametric Empirical Bayes (PEB) in that the prior distribution of the penultimate hierarchy level: $\pi_{L-1}(\eta_{L-1} \mid \eta_L)$ takes on a parametric form. Therefore, once the estimate of η_L is obtained from the data, the posterior distribution of θ_i is completely specified. Morris (1983) provides an excellent review of this PEB method. Efron (1996) provides discussion about using the Empirical Bayes method to combine likelihoods. Carlin and Louis (2000) summarize the past, present and future of the EB method.

Despite the above mentioned disadvantages the EB method has gained increased popularity in various application fields. Wong and Mason (1985) study the use of modern contraceptive methods in developing countries using the Empirical Bayes estimation procedure, Belin et al. (1993) studies undercount estimation problem in census data. Kahn and Raftery (1996) apply the EB method in medical research, to name a few areas of application.

Another approach, called the Non-parametric Empirical Bayes (NPEB) method, assumes that the penultimate hierarchy level distribution has an unknown form $g(\eta_{L-1})$. For more detailed discussion about NPEB, see Carlin and Louis (1996).

1.3.3.2 Fully Bayesian method

The Fully Bayesian method remedies the disadvantages of variance underestimation found in the EB method. It treats the hyper-parameter η_L as unknown with its own prior distribution $\pi(\eta_L)$ .When historical or expert information is available, we use an informative prior on the η_L , otherwise, a non-informative prior is utilized. Berger (1985) discusses various approaches to prior specification. Kass and Wasserman (1996) give a thorough account of prior selection. Since a non-informative prior is often improper ($\int \pi(\phi)d(\phi) = \infty$), it could lead to improper posterior distributions, in which case, all inference drawn from this posterior distribution is invalid. Therefore, when a non-informative prior is used, we must check that the posterior distribution is proper. A sensitivity study is generally recommended when a non-informative prior is used too see if the model is robust to changes of prior specification.

1.3.4 Small area estimation

One important application of hierarchical models is called Small Area Estimation. "Small areas" refer to small geographical regions or subpopulations under study with relatively small sample size. It becomes difficult to obtain estimates of interesting characteristics of certain subpopulations with adequate precision using information from that area alone. The distinctive "pooling strength" feature of hierarchical modeling makes use of population-level or neighborhood information to draw more valid inference about that small area with improved precision, which in our view and many other statisticians, is one of the most appealing aspects of hierarchical modeling idea. This kind of problem very often occurs in survey studies. For example, a nationwide survey about the certain disease occurrence is conducted with the aim of understanding the distribution of the occurrence at the national and state levels. Some policy makers, however, are particularly interested in estimating the mean, variance and the distribution at the county level. Some counties may be represented by too few (or no) occurrences due to the sampling process or some non-sampling error. Analysis based solely on that county would be unreliable

and combining information from the whole nation or from certain states usually results in more sensible estimates. One of the early uses of this technique can be traced back to Fay and Herriot (1979) in which U.S census income data are used to estimate income for small areas. Datta and Ghosh (1991) illustrate the use of the hierarchical linear model with application in small area estimation. Wong and Mason (1985) employ hierarchical models to analyze data from the World Fertility Survey. Ghosh and Rao (1994) review several techniques for dealing with small area estimation problem and illustrate many advantages of the empirical hierarchical Bayes method.

1.3.5 Bayesian computation

As stated before, the posterior distribution is computed as

$$p(\theta \mid y) = \frac{\int \ldots \int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)\ldots h(\eta_L)d\eta_1 d\eta_2 \ldots d\eta_L}{\int \ldots \int f(y \mid \theta)\pi_1(\theta \mid \eta_1)\pi_2(\eta_1 \mid \eta_2)\ldots h(\eta_L)d\theta d\eta_1 d\eta_2 \ldots d\eta_L}.$$

When the priors are not conjugate, the integral cannot be carried out analytically. Researchers used many approximation methods to tackle this problem, including the use of EM (Expectation- Maximization) algorithm and its variants (e.g. ECM, ECME), the Laplace method, etc. Even for moderately complex models, it could be very laborious to carry out the approximation. This computational difficulty has hindered the application of Bayesian methodology for almost two hundred years. Fortunately, Monte Carlo (MC) based sampling methods have been successfully developed over the past two decades, making the wide use of Bayesian methods possible Those methods include (but are not limited to) importance sampling (Ripley 1987), the Gibbs sampler (Gelfand and Smith, 1990) and Metropolis-Hastings sampling (Metropolis 1953, Hastings 1970, Gilks et al., 1996). The idea is based on the observation that anything one wants to know about a probability distribution can be learnt to arbitrary accuracy by sampling from it (Metropolis and Ulam, 1949). To draw valid inference from the sampled posterior distribution, it requires that the sampling process converges to the equilibrium or stationary distribution, which, unfortunately, is not always achievable. Thus, while the Fully Bayesian method is superior to the Empirical Bayes method from a theoretical

point of view, the use of non-informative priors or diffuse priors always makes the computation more challenging for the Fully Bayesian method than the Empirical Bayes method in that it is more difficult for the Fully Bayesian model to converge in a limited number of simulation iterations. That is one reason why the Empirical Bayes method is favored by some practitioners despite of some theoretical disadvantages.

1.3.5.1 Introduction of Markov chain Monte Carlo

We introduced MCMC methods briefly in the previous paragraphs. A key reference is the book by Gilks and Spiegelhalter (1996).

MCMC (Markov chain Monte Carlo) is essentially Monte Carlo integration using Markov chains. Markov chain simulation is based on a simulated random walk in the space of θ that converges to a stationary distribution, which is the desired or "target" distribution (in a Bayesian setting this is usually the posterior distribution: $p(\theta \mid y)$). The key to Markov chain simulation is to create a chain whose stationary distribution is exactly the target distribution and to run the simulation long enough so that the distribution of current samples is close to this stationary distribution.

The validity of inference based on MCMC simulations relies crucially on the convergence of the simulation process. As stated in the above theorem, when n approaches infinity, the ergodic property of the Markov chain ensures that the stationary distribution is exactly the target distribution. In reality, n can only be finite, therefore, we must ensure that the current distribution of samples is close to the stationary distribution. In other words, convergence must be established before inference is made. At least 10 convergence diagnostics have been proposed (Cowles and Carlin, 1994). The two most popular methods are proposed by Raftery and Lewis (1992) and Gelman and Rubin (1992). A problem with using a single chain to detect convergence is that a single chain may seem to have converged, but it is not when compared to several other chains with different starting values. It is possible that a single chain may seem to have converged perfectly when the simulation moves too slowly or the simulation gets stuck in certain

areas of the target distribution. In either case, convergence is clearly not established. Because of this potential risk with using a single chain, we use multiple chains to check convergence and the potential scale reduction factor (psrf) criteria proposed by Gelman and Rubin (1992) is used to assess convergence in our studies.

1.3.5.2 Gibbs sampler

The Gibbs sampler is one of the most popular MCMC sampling algorithms in Bayesian computation. It was formally introduced by Geman and Geman (1984) in a discussion of image restoration. Gelfand and Smith (1990) discussed its application in a variety of statistical problems.

Let Y be the data vector and $\theta = (\theta_1, \theta_2, ...\theta_p)$ be the parameter vector, $f(Y \mid \theta)$ the likelihood function and $\pi(\theta)$ the prior distribution. $\pi(\theta \mid Y) \propto f(Y \mid \theta)\pi(\theta)$ is the posterior distribution. $\pi(\theta_i \mid Y, \theta_{-i})$ is called the full conditional distribution of θ_i conditional on all other unknown θ's and the data.

The Gibbs sampler employs full conditional distributions to do the simulation. Full conditional distributions are more often available in closed form compared to marginal posterior distributions and therefore it is easier and more efficient to sample from these.

The Gibbs Sampler starts with a set of initial values of θ, namely $\theta_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0p})$ and updates the full conditional distributions as follows:

$$\theta_{i1} \sim f(\theta_1 \mid Y, \theta_{i-1,2}, \theta_{i-1,3}, ..., \theta_{i-1,p}),$$
$$\theta_{i2} \sim f(\theta_2 \mid Y, \theta_{i,1}, \theta_{i-1,3}, ..., \theta_{i-1,p}),$$
$$...$$
$$\theta_{ik} \sim f(\theta_k \mid Y, \theta_{i,1}, \theta_{i,2}, ..., \theta_{i,k-1,} \theta_{i-1,k+1}, ..., \theta_{i-1,p}),$$
$$...$$
$$\theta_{ip} \sim f(\theta_p \mid Y, \theta_{i,1}, \theta_{i,2}, ..., \theta_{i,p-1}),$$

The above procedure results in a single iteration of Gibbs sampler and produces $\theta_i=(\theta_{i1,}\theta_{i2,}...,\theta_{ip})$. This process is repeated *M* times. The initial *m* samples are usually thrown away to reduce the effect of the initial values. *M* should be large enough that the (*M-m)* samples of θ converge to a stationary distribution which is the target posterior distribution $\pi(\theta|Y)$ and all inference are then made on the samples obtained. The full conditionals change after every iteration since the conditioned values change for each iteration. Therefore, it is practically very important that sampling from the conditionals is highly computationally efficient. When the closed form of the full conditionals are not available in complex modeling situations, special techniques are needed to evaluate the full conditional function first before drawing samples from it. Possible techniques include rejection sampling, the ratio-of-uniform method, the Metropolis-Hastings algorithm and their variants and hybrids.

1.3.5.3 Metropolis-Hastings algorithm

We introduce the form by Hastings (1970), which is a generalization of the methods proposed by Metropolis et al. (1953). Let $\pi(.)$ be the target distribution which is generally not available in closed form, and $q(.|.)$ be the proposal distribution from which to draw samples. The algorithm works in the following way: at each iteration *t*,

1) Sample a point *T* from $q(.|X_t)$,

2) Sample *U* from the uniform distribution: *unif*(0,1),

3) If $U \le \min(1, \frac{\pi(T)q(X_t|T)}{\pi(X_t)q(T|X_t)})$, set $X_{t+1}=T$, otherwise, set $X_{t+1}=X_t$ (which means the chain does not move).

This process is repeated *M* times until convergence is achieved. Then $\{X_1,...,X_M\}$ form the sample for $\pi(.)$.

Theoretically $q(.|.)$ can be of any form as long as the chain will satisfy these three properties: irreducible, aperiodic and positive recurrent. Then the stationary distribution of the chain will be $\pi(.)$. However, the rate of convergence depends on the relationship

between $q(.|.)$ and $\pi(.)$. So in practice, it is worthwhile to do some exploratory analysis on the target distribution which will help construct a proposal distribution so the iterations will mix rapidly. If $\pi(.)$ is unimodal and not heavily tailed, a convenient proposal function of $q(.|.)$ could be a normal distribution whose location and scale is chosen to match $\pi(.)$. For more complex $\pi(.)$, $q(.|.)$ could be mixtures of normals or *t*-distributions. In addition, this $q(.|.)$ should be easy to sample from and evaluate to increase computation efficiency. Some special techniques include the Metropolis algorithm, independence sampler, single-component Metropolis Hastings and the aforementioned Gibbs sampler. For further details, see Robert (1995).

1.4 Motivating data

1.4.1 Brook trout

Brook trout (*Salvelinus Fontinalis*) is a type of fish native to northern North America. In the 1600s the spatial extent of brook trout (*Salvelinus Fontinalis*) in the eastern United States ranged from Georgia to Maine. In recent years, a large scale study was carried out to investigate the distribution, status and threats to brook trout in the eastern United States.

The study is described in detail in Hudy et al. (2006) and Thieling (2006). According to Hudy et al. (2006), anthropogenic physical, chemical and biological perturbations have resulted in a brook trout population decline in 59% of the subwatersheds in the eastern United States, which has raised concern from numerous state and federal agencies, non-government organizations and anglers. Understanding the relationships between brook trout population status and potential stressors is essential in developing useful managerial strategies for watershed level restoration, inventory and monitoring.

1.4.2 Data

The original data contains information about brook trout population status which falls into 7 categories. Sixty-three anthropogenic and landscape variables are compiled. They further combine subcategories into two big categories in terms of population status: extirpated or present. Extirpated regions are those in which brook trout were once present but are not longer considered present. In the remainder of this dissertation, we will use this binary response data as both motivation and example for our study.

The study area covers 16 states stretching from Maine to Georgia with complete data on 3337 subwatersheds. See Table 1.2.

Table 1.2: Brook trout data summary for each state

State	Sample size	Extirpated	Present
NEW HAMPSHIRE	47	0	47
VERMONT	186	6	180
MAINE	315	5	310
CONNECTICUT	175	29	146
MASSACHUSETTS	130	20	110
NEW YORK	350	115	235
PENNSYLVANIA	1085	444	641
NEW JERSEY	58	31	27
OHIO	4	1	3
MARYLAND	132	82	50
VIRGINIA	319	148	171
WEST VIRGINIA	174	24	150
SOUTH CAROLINA	19	12	7
NORTH CAROLINA	214	95	119
TENNESSEE	54	18	36
GEORGIA	75	53	22
TOTAL	**3337**	**1083**	**2254**

The pattern of extirpation varies considerably over the large spatial region. A single model approach, though convenient to use, may not be appropriate to accurately describe the relationship between the response variable and the explanatory variables. Hudy et al. (2006) investigate a variety of modeling approaches for predicting extirpation of brook trout treating the entire data as if they come from the same homogeneous distribution. Using classification trees, they were able to develop a model that produced reasonably good prediction with five potentially causative variables. While correct classification rates were good for these models, there were several regions where classification rates were observed to be relatively weak.

We propose an alternative modeling approach with the goal of better classification performance.

Chapter 2 Bayesian Hierarchical Logistic Model

In environmental monitoring, binary response data are encountered often. For example, the water quality of certain water body may be classified as "impaired" or "not impaired" based on certain standards. Certain species can be either "present" or "absent" in certain streams. The logistic model is a commonly used regression model to relate the binary response to the explanatory variables. It has good properties with respect to asymptotic efficiency over other non-parametric or semi-parametric approaches in dealing with classification problem. When data display hierarchical structure, Bayesian hierarchical modeling provides a better alternative to either a grand model or several separate models for estimating location-specific parameters while taking into account the correlations among the data. Using brook trout data with a binary response to indicate the condition of the streams under study, we developed a predictive hierarchical logistic model for classifying brook trout population status in certain streams. Empirical Bayesian methods were used in developing the prior distributions and the Markov chain Monte Carlo (MCMC) method was used in obtaining posterior inference. Furthermore, predictive performance of both non-hierarchical logistic models and hierarchical logistic models were compared. The result shows that the Bayesian hierarchical model outperforms the non-hierarchical model.

2. 1 Introduction

In water quality monitoring, studies are rarely implemented in isolation due to the geographical distribution of the water systems involved in such studies. This leads to a

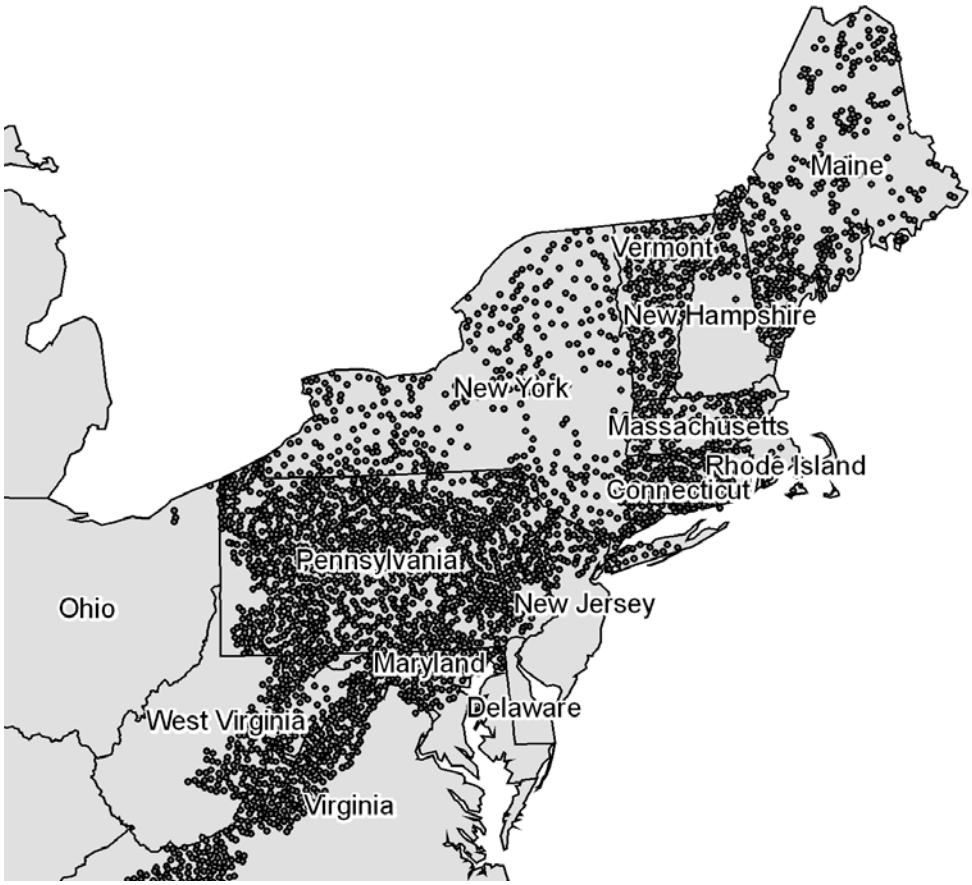
hierarchical structure of the data. In such cases, parameters governing each hierarchy level are usually not independent due to the hierarchical structure and therefore a joint probability distribution is assumed upon those parameters to reflect their dependence. This is the basic idea of Bayesian hierarchical modeling (Kass and Steffey, 1989). Such modeling strategy, either from a fully or empirical Bayesian perspective, helps deal with the multi-parameter problem (multiple cluster-specific parameters involved in a study) , usually improves precision of each individual parameter estimate through "pooling strength" and performs better in terms of predictive ability (Gelman 1995, Congdon 2001, Congdon 2003). For this reason, Bayesian hierarchical modeling has seen successful application in many fields. For modeling binary and/or multinomial data, researchers have been integrating logistic regression into the Bayesian hierarchical model framework. Wong and Mason (1985) developed such model in studying the use of modern contraceptive methods in developing countries. Wolpert and Warren-Hicks (1992) applied this model in analyzing ecological survival data. Belin et al (1993) studied undercount estimation problem in census data. Kahn and Raftery (1996) used this idea in medical research areas. Gelman and Little (1997) employed this method to study U.S election poll data. Scott and Ip (2002), Browne and Draper (2004) applied this method in educational studies.

The project we are interested in was initially carried out by Hudy et al. (2006) and Thieling (2006) in which the population status of brook trout in streams is investigated across 16 eastern states in the United States. Different regions may have different environmental variation affecting brook trout population differently. We could fit a grand model to all the data points ignoring the spatial correlation among them, which may lead to poor fit of the data. We could fit separate models to different locations, which usually results in better fit of the current data but poor prediction of future observations and perhaps over optimistic views. To overcome both disadvantages, we present a Bayesian hierarchical logistic modeling approach which serves as a compromise between these two single-level approaches while taking into account both the similarities and differences among locations.

2. 2 Background

A recent study was carried out recently on the distribution, status and threats on brook trout in the eastern United States (Thieling, 2006). Brook trout (*Salvelinus Fontinalis*) are the only trout native to many of the eastern United States. Brook trout are a good source of table food. They also serve as indicators of the health of the watersheds they inhabit (Trout Unlimited, 2006). A strong brook trout population indicates that water quality in a stream or river is excellent. A decline in brook trout population indicates that a stream or river is at risk. The original data contain information about 63 anthropogenic and landscape variables and the response variable: brook trout population status, which falls into 7 categories. They further combine the 7 subcategories into two categories in terms of population status: extirpated or present. Extirpated regions are those in which brook trout were once present but are not longer considered present. See Figure 2.1 for the geographical layout of the study area.

Figure 2.1: Geographical layout of the brook trout study area. (*N*=3337)



Most of the 63 variables are percentage variables. We apply the Box-Cox transformation approach and either logarithm transformation or square root transformation is used on some of the original variables to linearize the metrics and to offset the effects of some outliers in the dataset.

2. 3 Variable screening

There are a large number of variables in the original data. We performed the following variable screening to select the important ones to build a classification model. We did the following analyses:

- Correlation analysis. Since many of the variables are correlated, we performed a correlation analysis first. For a pair of correlated variables (correlation ≥ 0.8), only one variable will be kept for further screening.
- P-value calculation. A logistic regression model was fit and the partial p-value for each variable in the logistic regression model was calculated and ranked.
- Stepwise selection. A stepwise variable selection procedure was carried out when fitting the variables to a logistic regression model.
- Classification Trees analysis. We also fit those variables into a classification tree program looking for significant variables.

We carefully reviewed all the numerical results obtained from the above screening process. We also consulted field experts and obtained suggestions regarding the practical importance of certain variables. We decided to use the following five metrics to build the final predictive model. They are Elevation (mean elevation of subwatershed), Acid deposition (combined concentration of nitrate (NO_3) and sulfate (SO_4)), Road density (subwatershed road density, km of road per km^2 of land), Agriculture (percentage of subwatershed agricultural use) and Total forest (percentage of subwatershed forested lands). Transformations are applied to four of these variables. See Table 2.1 for the transformation details. We further center and scale these variables. Centering and scaling has been long recognized as a proper

procedure in regression analysis (Myers, 1990). It also helps implementing the iterative computation of logistic models.

Table 2.1: Transformation applied to the predictor variables

Variable name	**Transformation**
Elevation	None
Acid deposition	natural logarithm
Road density	natural logarithm
Agriculture	square root
Total forest	square root

2. 4 Hierarchical structure of the data

The study area of this brook trout project ranges from Georgia to Maine covering 16 eastern states of the United States as showed in Figure 2.1 above. The pattern of extirpation varies considerably over the large spatial region. With the aid from some fishery experts, namely, Mark Hudy and his colleagues at James Madison University, we grouped certain neighboring states into subregions forming an empirical hierarchical structure of the data as summarized in Table 2.2. It is assumed that data points within the same region are somewhat similar to each other while different across different regions.

Table 2.2: Hierarchical structure of the brook trout data

Region	State	Total	Extirpated	Present
1	NEW HAMPSHIRE	47	0	47
	VERMONT	186	6	180
	MAINE	315	5	310
	Subtotal	**548**	**11**	**537**
2	CONNECTICUT	175	29	146
	MASSACHUSETTS	130	20	110
	NEW YORK	350	115	235
	Subtotal	**655**	**164**	**491**
3	PENNSYLVANIA	1085	444	641
	NEW JERSEY	58	31	27
	OHIO	4	1	3
	Subtotal	**1147**	**476**	**671**
4	MARYLAND	132	82	50
	VIRGINIA	319	148	171
	WEST VIRGINIA	174	24	150
	Subtotal	**625**	**254**	**371**
5	SOUTH CAROLINA	19	12	7
	NORTH CAROLINA	214	95	119
	TENNESSEE	54	18	36
	GEORGIA	75	53	22
	Subtotal	**362**	**178**	**184**

2. 5 Model

We set up the hierarchical model in the following way:

$$
\begin{aligned}
&Y_{ij} \mid p_{ij} \sim Bernoulli\,(p_{ij}) \\
&\text{logit}(p_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}_i \\
&\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \Sigma \sim N_k(\boldsymbol{\mu}, \Sigma^{-1}) \\
&\pi(\boldsymbol{\mu}) \sim N_k(\boldsymbol{\alpha}, \Omega) \\
&\pi(\Sigma) \sim Wishart(\mathrm{R}, v)
\end{aligned}
$$

where $i = 1,2,...,5$ indexes regions and $j = 1,2,...,n_i$ indexes observations of a region. The *j*th observation in the *i*th region Y_{ij} is assumed to follow a Bernoulli distribution with event probability p_{ij}. A logistic regression model is used to model p_{ij} with the logit link and the linear predictor to be $\mathbf{x}_{ij}'\boldsymbol{\beta}_i$. The $\boldsymbol{\beta}_i$'s are the parameter vector for each region. The dimension of $\boldsymbol{\beta}_i$ is denoted by *k* and is 6 in this case (we have five predictor variables plus an intercept term in the model). Since all five regions are in the eastern United States, they share certain common geographical features. Therefore, it is reasonable to assume a prior distribution on for all the $\boldsymbol{\beta}_i$ to incorporate the connections between study regions rather than model them separately as if they are independent to each other. Specifically, the $\boldsymbol{\beta}_i$'s are assumed to come from the same multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and precision matrix $\sum$. The hyper prior distribution for $\boldsymbol{\mu}$ and $\sum$ are the commonly used multivariate normal distribution and Wishart distribution respectively. Three hyper-parameters are involved in this setup. They are: $\boldsymbol{\alpha}$, the mean vector of the hyper-prior distribution for $\boldsymbol{\mu}$; Ω, the variance-covariance matrix of the hyper-prior distribution for $\boldsymbol{\mu}$ and R, the variance matrix of the prior distribution for the precision matrix $\sum$. The empirical estimates of the parameter vector and covariance matrix are obtained from a logistic regression model on all the data. The estimated parameter vector is used as the value for $\boldsymbol{\alpha}$. The estimated covariance matrix could be used for Ω and R. Based on our simulation result, we decide to use identity matrix for both Ω and R. This way, the variance matrix is less restrictive than the crude estimate but not too vague to cause convergence problem. A non-informative Wishart distribution is obtained as the degree of freedom $\nu \to 0$. The density is finite if $v \geq k+1$.Therefore, *v* is set to be 7 (dimension of R matrix plus 1) which is the minimum value to make the density finite.

2. 6 Implementation

Due to the complexity of the model, the posterior distributions of the parameters of interest are not available in closed form. The Markov chain Monte Carlo (MCMC)

method is used to implement this logistic hierarchical model. In particular, a hybrid of Gibbs sampling (Gelfand and Smith, 1990) and the Metropolis algorithm (Metropolis, 1949, 1952) is used to obtain the posterior distributions of parameters of interest.

Since the marginal posterior distribution of the parameters of not available analytically, we compute the full conditional distribution of all parameters. We first obtain the joint distribution of Y, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and Σ:

$$\prod_{i=1}^{m}\{\prod_{j=1}^{n_i} f(Y_{ij} \mid \boldsymbol{\beta}_i)\cdot \pi(\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \Sigma)\}\cdot \pi(\boldsymbol{\mu})\cdot \pi(\Sigma)$$

$$= \prod_{i=1}^{m}\left(\prod_{j=1}^{n_i}\left(\frac{\exp(\mathbf{x}_{ij}'\underline{\boldsymbol{\beta}}_i)}{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_i)}\right)^{y_{ij}}\left(\frac{1}{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_i}\right)^{1-y_{ij}} \cdot \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta}_i-\boldsymbol{\mu})'\Sigma(\boldsymbol{\beta}_i-\boldsymbol{\mu})\right)}{\sqrt{2\pi}^{k}\mid \Sigma^{-1}\mid^{1/2}}\right)\cdot$$

$$\cdot \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\alpha})'\Omega^{-1}(\boldsymbol{\mu}-\boldsymbol{\alpha})\right)}{\sqrt{2\pi}^{k}\mid \Omega\mid^{1/2}}\cdot \frac{\mid \Sigma\mid^{\frac{v-k+1}{2}}\cdot \mid \mathrm{R}\mid^{-\frac{v}{2}}\cdot \exp\left(-\frac{1}{2}trace(\mathrm{R}^{-1}\Sigma)\right)}{2^{\frac{vk}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma\left(\frac{v+1-i}{2}\right)}$$

where *m* is the number of groups, in our case, *m*=5 indicating the five regions under study.

The prior distributions for $\boldsymbol{\beta}_i$'s are not conjugate and the posterior distributions are therefore not available in closed form. The full conditional for $\boldsymbol{\beta}_i$ can be written as

$$\pi(\boldsymbol{\beta}_i \mid \cdot) \propto \prod_{j=1}^{n_i} f(Y_{ij} \mid \boldsymbol{\beta}_i)\cdot \pi(\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\propto \prod_{j=1}^{n_i}\left(\frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_i)}{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_i)}\right)^{y_{ij}}\cdot\left(\frac{1}{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_i)}\right)^{1-y_{ij}}\cdot \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_i-\boldsymbol{\mu})'\Sigma(\boldsymbol{\beta}_i-\boldsymbol{\mu})\right)$$

which is not available in closed form either.

The full conditional distribution of $\boldsymbol{\mu}$ can be written as

$$\pi(\boldsymbol{\mu}\mid\cdot)\propto\prod_{i=1}^{m}\pi(\boldsymbol{\beta}_i\mid\boldsymbol{\mu},\Sigma)\pi(\boldsymbol{\mu})$$
$$\propto\exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\mu}-\boldsymbol{\beta}_i)'\Sigma(\boldsymbol{\mu}-\boldsymbol{\beta}_i)\right)\cdot\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\alpha})'\Omega(\boldsymbol{\mu}-\boldsymbol{\alpha})\right)$$
$$\propto\exp\left(-\frac{1}{2}(\boldsymbol{\mu}'(m\Sigma+\Omega^{-1})\boldsymbol{\mu}-2(\sum_{i=1}^{m}\boldsymbol{\beta}_i\Sigma+\boldsymbol{\alpha}'\Omega^{-1})\boldsymbol{\mu})\right)$$
$$\propto\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-(m\Sigma+\Omega^{-1})^{-1}\cdot(\sum_{i=1}^{m}\boldsymbol{\beta}_i\Sigma+\boldsymbol{\alpha}'\Omega^{-1})')'\cdot(m\Sigma+\Omega^{-1})\cdot(\boldsymbol{\mu}-(m\Sigma+\Omega^{-1})^{-1}\cdot(\sum_{i=1}^{m}\boldsymbol{\beta}_i\Sigma+\boldsymbol{\alpha}'\Omega^{-1})')\right)$$

which follows a multivariate normal distribution:

$$MVN\left((m\Sigma+\Omega^{-1})^{-1}\cdot(\sum_{i=1}^{m}\boldsymbol{\beta}_i\Sigma+\boldsymbol{\alpha}'\Omega^{-1})',(m\Sigma+\Omega^{-1})^{-1}\right)$$

The full conditional distribution of $\sum$ can be written as

$$\pi(\Sigma\mid\cdot)\propto\prod_{i=1}^{m}\pi(\boldsymbol{\beta}_i\mid\boldsymbol{\mu},\Sigma)\cdot\pi(\Sigma)$$
$$\propto|\Sigma|^{\frac{m}{2}}\cdot\exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\mu}-\boldsymbol{\beta}_i)'\Sigma(\boldsymbol{\mu}-\boldsymbol{\beta}_i)\right)\cdot|\Sigma|^{\frac{v-k+1}{2}}\cdot\exp\left(-\frac{1}{2}trace\,(\mathrm{R}^{-1}\Sigma)\right)$$
$$\propto|\Sigma|^{\frac{v+m-k-1}{2}}\cdot\exp\left(-\frac{1}{2}\left(trace\left(\mathrm{R}^{-1}\Sigma\right)+trace\left(\sum_{i=1}^{m}(\boldsymbol{\mu}-\boldsymbol{\beta}_i)(\boldsymbol{\mu}-\boldsymbol{\beta}_i)'\Sigma\right)\right)\right)$$
$$\propto|\Sigma|^{\frac{v+m-k-1}{2}}\cdot\exp\left(-\frac{1}{2}trace\left(\left(\mathrm{R}^{-1}+\sum_{i=1}^{m}(\boldsymbol{\mu}-\boldsymbol{\beta}_i)(\boldsymbol{\mu}-\boldsymbol{\beta}_i)'\right)\Sigma\right)\right)$$

which follows a Wishart distribution:

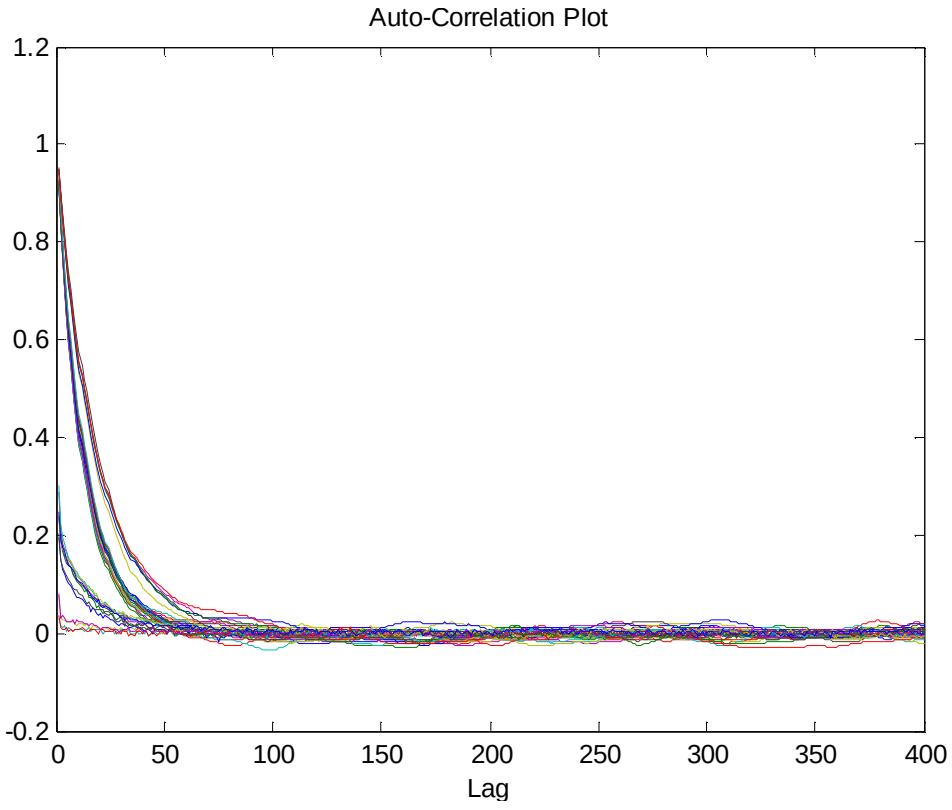
$$Wishart\left(\left(\mathrm{R}^{-1}+\sum_{i=1}^{m}(\boldsymbol{\mu}-\boldsymbol{\beta}_i)(\boldsymbol{\mu}-\boldsymbol{\beta}_i)'\right)^{-1},v+m\right)$$

The full conditional distributions for $\boldsymbol{\mu}$ and $\sum$ are known distributions. Gibbs sampling can then be used to simulate samples from those full conditional distributions. The full conditional distribution for $\boldsymbol{\beta}_i$ is not tractable, thus, we employ Metropolis algorithm to do the sampling using a multivariate normal distribution as the proposal distribution. See Appendix A for more computational details.

The computation is implemented in MATLAB. Two chains of simulations are used with each run of size 105,000. For each chain, a burn-in of 5000 runs is used to eliminate the

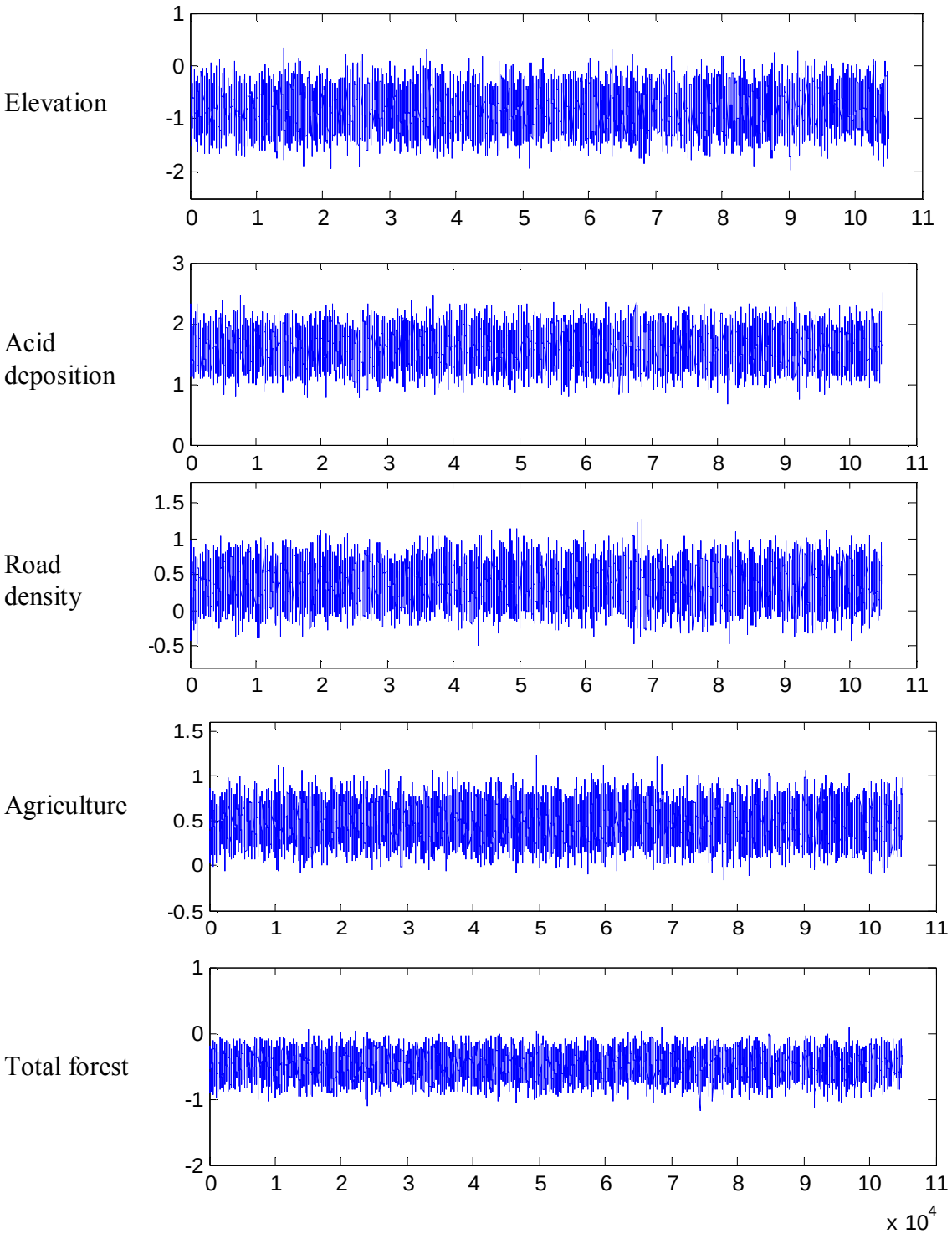
effect of starting values and a thinning of 50 is used to ensure that the auto-correlation drops to close to zero. Graphical diagnostics including auto-correlation plot and trace plots (see Figure 2.2 and Figure 2.3 demonstration) shows the simulation process has converged. The more formal diagnostic tools: the Potential Scale Reduction Factor (PSRF) (Gelman and Rubin, 1992) and the Multivariate Potential Scale Reduction Factor (MPSRF) (Brooks and Gelman, 1998) also suggest convergence. The mpsrf values for $\boldsymbol{\beta}_i$'s, $\boldsymbol{\mu}$ and $\sum$ are listed in Table 2.3. The final analysis is based on 4000 samples combined from two chains after burn-in and thinning.

Figure 2.2: Auto-correlation plot for the five $\boldsymbol{\beta}_i$ vectors from chain 1



Note: Plot from chain 2 is similar and is omitted

Figure 2.3: Trace plots of the coefficients for the 5 predictor variables for region 2



Note: The variable name is indicated on the left side of each plot. The y-axis is the simulated value of the coefficient and the x-axis is the iteration number. The rest of the plots show similar patterns and are therefore omitted.

Table 2.3: Multivariate Potential Scale Reduction Factor

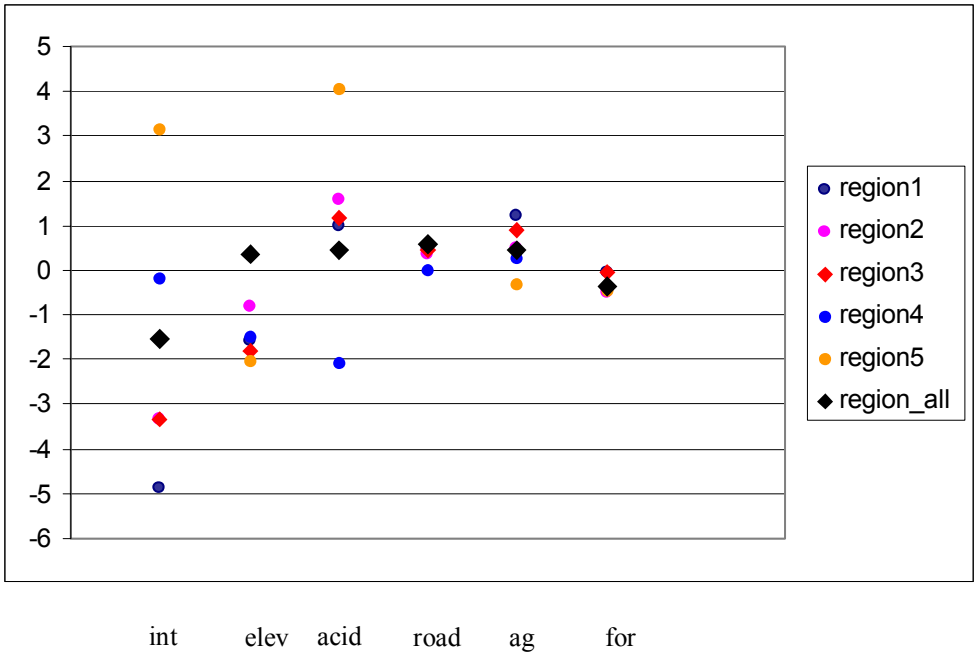
MPSRF	β_1	β_2	β_3	β_4	β_5	μ	Σ
Parameter	1.0008	1.0006	1.0025	1.0005	1.0012	1.0002	1.005

Note: Values below 1.1 indicate convergence.

2. 7 Coefficients estimates

We obtain estimates (specifically, posterior means) for each β_i using the Bayesian hierarchical logistic modeling approach. The results are summarized in Table 2.4. We also apply the non-hierarchical logistic model on all the data combined. The Maximum likelihood estimates are summarized in Table 2.5. The Bayesian estimates for each region and the maximum likelihood estimates for the entire region are plotted in Figure 2.4.

Figure 2.4: Plots of region-specific coefficient estimates for the hierarchical model and the ML estimates for the entire region using the non-hierarchical model



Note: int--------intercept elev-------elevation
acid-------acid deposition road-------road density
ag---------agriculture forest-----total forest

Table 2.4: Regression coefficients estimates for the 5 regions from the Bayesian hierarchical logistic modeling approach

	Region				
	1	**2**	**3**	**4**	**5**
Variable	**mean (sd)**	**mean (sd)**	**mean (sd)**	**mean (sd)**	**mean (sd)**
Intercept	-4.88 (0.74)	-3.34 (0.31)	-3.34 (0.48)	-0.17 (0.19)	3.13 (0.46)
Elevation	-1.60 (0.54)	-0.83 (0.30)	-1.79 (0.25)	-1.49 (0.24)	-2.03 (0.25)
Acid Deposition	1.01 (0.46)	1.59 (0.23)	1.17 (0.43)	-2.07 (0.33)	4.06 (1.13)
Road density	0.48 (0.40)	0.36 (0.22)	0.45 (0.12)	-0.02 (0.17)	0.58 (0.20)
Agriculture	1.21 (0.48)	0.48 (0.17)	0.89 (0.13)	0.26 (0.23)	-0.32 (0.32)
Total forest	-0.07 (0.35)	-0.49 (0.16)	-0.03 (0.13)	-0.41 (0.23)	-0.45 (0.45)

Note: mean-------posterior mean sd------standard deviation

Table 2.5: Maximum likelihood estimates of the regression coefficients from the non-hierarchical logistic model using all the data

All regions	
Variable	**Estimate (s.e)**
Intercept	-1.536 (0.061)
Elevation	0.338 (0.058)
Acid Deposition	0.468 (0.069)
Road density	0.577 (0.062)
Agriculture	0.468 (0.060)
Total forest	-0.358 (0.065)

Note: s.e------standard error

It can be seen that the signs, magnitude and significance of the coefficient estimates vary across the 5 regions in the hierarchical modeling approach. For example, intercept reflects baseline conditions to some extent. The estimated intercept for region 5 is positive 3.134, which is dramatically different from the negative estimates (in the neighborhood of -4) for region 1, 2 and 3. The estimated coefficients for Acid deposition range from -2.070 to 4.059. This variability is quite large considering the variables are all centered and scaled. Discussions of other similar observations 5 are omitted. Overall, each region shows certain degrees of dissimilarities. Therefore the use of the hierarchical model is confirmed since the differences are allowed in individual estimates while the dependence among them is naturally taken into consideration by the hierarchical setup of the model.

From the non-hierarchical modeling approach, the coefficient estimate for elevation is positive 0.338. Elevation is partially an indicator of the temperatures in the watersheds. Since we model the probability of brook trout being extirpated from certain streams, the positive estimate of elevation suggests that the colder the place is (with higher elevation), the more likely brook trout will be extirpated. This is contradictory to the fact that brook trout have a preference for cooler streams. When we look at the estimates obtained from the hierarchical modeling approach, elevation has a uniformly negative effect on brook trout being extirpated. This is more sensible and confirms experts' findings (Hudy, et al., 2006).

2. 8 Classification performance

We are most interested in the model's predictive ability. We compute the two measures for a model's predictive accuracy for both the hierarchical models and non-hierarchical models. They are Average Fraction of Correctly Classified for Fit (AFCCF) and the area under the ROC curve (AUC).

2.8.1 AFCCF

In the non-hierarchical model, $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}$,where $\hat{\boldsymbol{\beta}}$ is the MLE, is used as the predicted probability. In the Bayesian hierarchical model, the posterior distribution of $\boldsymbol{\beta}$ is obtained. The posterior mean and/or posterior median are commonly used as the point estimate for $\boldsymbol{\beta}$. We calculate AFCCF for the hierarchical model using $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}$. The result using the posterior median as the estimate for $\boldsymbol{\beta}$ is the same as using the posterior mean estimates keeping two decimal places and is omitted. The results for the AFCCF measure are 0.76 for the hierarchical modeling approach and 0.68 for the non-hierarchical modeling approach. The hierarchical model has an improvement of 12% over the non-hierarchical model.

2.8.2 AUC

We obtain the AUC values for both hierarchical and non-hierarchical modeling approaches. In the non-hierarchical model, $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}$, where $\hat{\boldsymbol{\beta}}$ is the MLE, is used as the predicted probability. In the Bayesian hierarchical model, $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}$ where $\hat{\boldsymbol{\beta}}$ is the posterior mean is used as the predicted probability in computing the AUC. The results for the AUC measure are 0.90 for the hierarchical modeling approach and 0.80 for the non-hierarchical modeling approach.

The gain from the hierarchical model is significant. It improves prediction for about 13% based on the AUC criterion. An AUC value of 0.90 is generally considered to indicate a model's excellent classification ability. Our hierarchical model is able to achieve this.

2.9 Concluding remarks

In this chapter, we studied the relationship between several environmental metrics and the population status of brook trout in the eastern United States. We developed a Bayesian hierarchical logistic model to account for the geographical correlation among several regions involved in this water quality monitoring effort. Predictive ability measures including AFCCF and AUC are calculated and are compared to those obtained through MLE using a non-hierarchical logistic model. The results indicate that the Bayesian hierarchical model outperforms the non-hierarchical logistic model. The success of our hierarchical model over the non-hierarchical model stems from the fact that the natural geographical hierarchy of data is taken into consideration such that both similarity and dissimilarity between clusters of observation are simultaneously modeled. In environmental studies, data with hierarchical structure due to the spatial correlation among observations are sometimes encountered. When applicable, we recommend the use of Bayesian hierarchical models. With successfully developed Monte Carlo (MC) based sampling methods over the last two decades (Gilks, Richardson, and Spiegelhalter, 1996), Bayesian hierarchical models in the environmental studies are conceptually favorable and computationally feasible.

Chapter 3 Model-Based Clustering

When dealing with data compiled over a large spatial region, a single model may not be appropriate to describe relationships between variables. We developed a model-based clustering method to group categorical response data by their empirical stressor-response relationships with the goal of better classification performance after clustering. Voronoi tessellation techniques are implemented to subdivide a region and the area under the receiver operator characteristic curve is used as the criterion when searching for the optimal clustering. This method is applied to a brook trout absence/presence data within subwatersheds of the eastern United States. Results indicate fairly strong stressor-response relationships that vary spatially and show significant improvement over the conventional single model approach.

3.1 Introduction

3.1.1 Background

In the 1600s the spatial extent of brook trout (*Salvelinus Fontinalis*) in the eastern United States ranged from Georgia to Maine. Human perturbations have lead to the extirpation of brook trout in many of the streams in which it once existed. To help manage brook trout populations and to prevent further loss of species occurrence, it is necessary to investigate possible causative factors and to model their influence on the loss of the species. Thieling (2006) investigates a variety of modeling approaches for predicting extirpation of brook trout using a data set from the eastern United States. Using classification trees, they were able to develop a model that produced reasonably good prediction with five potentially causative variables. While correct classification rates

were good for these models, there were several regions where classification rates were observed to be relatively weak, namely, the eastern Pennsylvania sites, some North Carolina, Georgia and Tennessee sites. This suggested that different models might be needed in these regions.

3.1.2 Single model vs. multiple models

Our concern in this chapter is the use of a single model to describe the relationship over a large region. When the spatial extent is quite large, it is reasonable to expect that the model may change or be different in different regions. For example, it might be anticipated that the effect of agricultural practices might be different in northern regions relative to southern ones. Effects of various stressors may be different in higher elevations than in lower ones. One approach used to account for location differences is to divide the whole dataset into clusters (e.g. ecoregions) such that data points are similar within each cluster and dissimilar among clusters, and then apply separate models accordingly. This approach is the basis of biological monitoring procedures such as RIVPACS or AUSRIVAS (Wright et al., 1984; Nichols et al., 2000). These procedures first cluster sites on natural factors that influence biological conditions, then model the relationships of biological responses with stressors. These methods are inadequate for our purposes since there is no guarantee that the models that result will have good correct classification rates. Making groups of sites as different as possible on natural factors may reduce relationships with stressors if the stressor gradient is connected in some way with the natural variables or if the stressor gradient crosses the cluster boundaries. What seems more sensible is to use a clustering procedure that is based on finding regions with strong relationships between extirpation and stressors.

3.1.3 Partitioning models

One approach is based on the idea of partitioning models. For example, Holmes, Denison and Mallick (1999) proposed a Bayesian Partitioning Model (BPM) that could be used to split a large region into disjoint sub-regions. Their examples use regression models that assume normal or multinomial responses. They have also extended the method to count data analysis (Denison and Holmes, 2001). Chipman, George and McCulloch (2002) proposed Bayesian Treed Models (BTM) as an extension of classification trees. The method uses a set of variables to split the data in a manner similar to what is done in classification trees but then uses a logistic or normal regression model as the final step in each node of the classification tree. It allows for richer models in each partition but permits only axis-parallel partitions, producing clusters of rectangular shape.

3.1.4 Overview of the model-based clustering method

The method we present is a variation of the approach of Prins et al. (unpublished document) and focuses on classification analysis using logistic models for prediction. We use one set of variables, latitude and longitude, to determine boundaries for clusters and a second set for modeling the relationship between the probability of extirpation and watershed stressor variables. To form clusters we use a randomization approach to divide the region into clusters. For a given number of clusters, k, random seeds are generated and used to create a Voronoi diagram. The diagram divides the space in polygonal regions. Within each region, a logistic regression model is used to model the stressor-response relationship. By repeating the process a large number of times with different random partitions, we are able to find clusters with the best or near best models for predicting extirpation at the subwatershed level.

To determine whether a particular clustering solution is optimal, we use a prediction based approach using the receiver-operator characteristic (ROC) curve. The logistic

model is fitted to data and a probability of extirpation is calculated. When this probability exceeds a critical value, the site is classified as extirpated. The ROC curve measures predictive ability over a range of critical probabilities. The area under the ROC curve (AUC) is used as the criterion in the Monte Carlo search for the optimal clustering solution. Cross validation is used to avoid bias from overfitting the number of clusters.

3.2 Motivating data

The dataset is described in detail in Hudy et al. (2006) and Thieling (2006) in which the distribution, status and threats to brook trout (*Salvelinus Fontinalis*) within the eastern United States are discussed. According to Hudy et al. (2006), anthropogenic physical, chemical and biological perturbations have resulted in a significant loss (>50%) of the self-sustaining brook trout streams in 59% of the subwatersheds in the eastern United States, raising concern from numerous state and federal agencies, non-government organizations and anglers. Understanding the relationships between brook trout population status and the perturbations is essential in developing useful managerial strategies for watershed level restoration, inventory and monitoring. More details can be found from Thieling (2006) or Hudy et al. (2006).

The study area covers 16 states stretching from Maine to Georgia with complete data on 3,337 subwatersheds. The candidate stressor metrics compromise 63 anthropogenic and landscape variables. The response variable we use is self-sustaining brook trout population status: extirpated (E) or present (P). Extirpated subwatersheds are those in which historic self-sustaining brook trout have all been lost. The pattern of extirpation varies considerably over the large spatial region (Table 3.1). We exclude 3 states in the study since the response is nearly uniform in those states (98% of the locations have brook trout present). This reduces the sample size to 2789, out of which 1717 are presence and 1072 are extirpated. We carefully screened the 63 candidate metrics based on redundancy (keeping only one variable for further screening for a pair of variables with high correlation) and significance to the response variable (using Wald chi-square test and classification tree method to search for significant variables). Four predictor

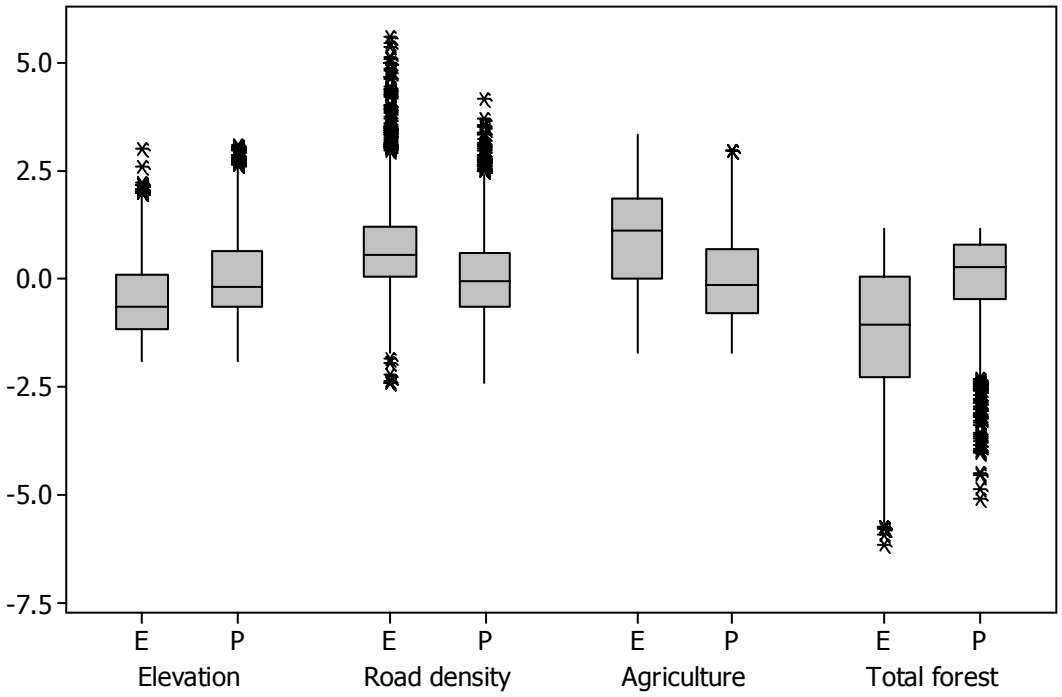
variables are chosen to build the final model. They are Elevation (mean elevation of subwatershed), Road density (subwatershed road density, km of road per km^2 of land), Agriculture (percentage of subwatershed agricultural use) and Total forest (percentage of subwatershed forested lands). We applied the Box-Cox transformation approach and the transformed variables as follows: Road density (logarithm transformation), Agriculture (square root transformation) and Total forest (square root transformation). We further center and scale all 4 variables using mean and standard deviation calculated from the reference (presence) group. Box-plots for the transformed variables are given in Figure 3.1.

Table 3.1: Data summary for each state

State	Sample size	Extirpated	Present
NEW HAMPSHIRE *	47	0	47
VERMONT *	186	6	180
MAINE *	315	5	310
Total	**548**	**11**	**537**
CONNECTICUT	175	29	146
MASSACHUSETTS	130	20	110
NEW YORK	350	115	235
PENNSYLVANIA	1085	444	641
NEW JERSEY	58	31	27
OHIO	4	1	3
MARYLAND	132	82	50
VIRGINIA	319	148	171
WEST VIRGINIA	174	24	150
SOUTH CAROLINA	19	12	7
NORTH CAROLINA	214	95	119
TENNESSEE	54	18	36
GEORGIA	75	53	22
Total	**2789**	**1072**	**1717**

* States not included in our model.

Figure 3.1: Box plots for the 4 transformed variables for the entire dataset



3.3 Method

Our approach to clustering models is based on three steps. First, a random set of points within the region are selected and used to partition the space. Second, within each partition, a logistic model is fitted. Third, the quality of the fit is evaluated. These three steps are repeated a large number of times, varying the number of partitions and the best value of the criterion is used to determine the final regions and models. Details are discussed below.

3.3.1 Formulation of k clusters

The first step involves partitioning the region into *k* clusters and this step involves use of Voronoi diagrams and Dirichlet tessellations (Møller, 1994, Okabe, et al., 2000). To do

this, we randomly generate k points (labeled $\mathbf{g}_1,...,\mathbf{g}_k$) within the region that form the spatial centers of the clusters then locate the group of points that are closest to the centers.

In the context of spatial data, a natural measure of closeness between sites i and j is the Euclidean distance based on the longitude and latitude i.e.

$$D(\mathbf{x_i}, \mathbf{x_j}) = \sum_{r=1}^{2} (x_{ir} - x_{jr})^2$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ represent the longitude and latitude for sites i and j, respectively. We calculate $D(\mathbf{x_i}, \mathbf{g_j})$ for all points ($\mathbf{x_i}$, i=1,2,…n) and cluster centers $\mathbf{g_j}$ (j=1,2,…k) and assign the point to the closest cluster.

3.3.2 Fitting regression models within clusters

Given a clustering of the points, we use a parametric model to fit the clustered data. The most commonly used model for categorical (binary) response data is logistic regression model with the event probability modeled by

$$\text{logit}(p) = \log(\frac{p}{1-p}) = x\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m,$$

where x_1 through x_m are the explanatory variables and $\boldsymbol{\beta}$ is a vector of $m+1$ regression parameters with the first parameter corresponding to the intercept term. Logistic regression models are useful for management purposes in that parameter estimates measure the importance of a variable given the other variables and may help in selection of management strategies. Statistically the model has good properties with respect to asymptotic efficiency over other non-parametric or semi-parametric approaches (Pepe, 2005) and is also quite robust against link violation (Li and Duan, 1989). Cluster specific multiple logistic regression models using the m predictor variables are fit to the data in the k clusters formed in step 1.

3.3.3 Calculating performance criteria measure

The third step is the choice of a measure of performance for the current clustering solution. The ROC curve is a graphic display that gives a measure of the predictive (discriminant) accuracy of a classification model. The area under the ROC curve (AUC) is the most common summary index describing an ROC curve and has long been used as a measure of classification performance (Bamber, 1975, Hanley and McNeil, 1982, Hanley, 1998, Pepe, 2000, 2003, 2005, Ma and Huang, 2005). The usual estimator for this area can be written as

$$\mathrm{AUC} = \frac{1}{n_E n_P} \sum_{i \in E, j \in P} I(p_i > p_j)$$

where E and P are the index sets for extirpated and present groups with size n_E and n_P respectively; p_i and p_j are the predicted probability of being extirpated for the *i*th (*j*th) observation in the extirpated (present) group and *I* corresponds to the indicator function (the indicator function assigns the value one if the condition is true). It is interesting to note that AUC is closely related to the Mann-Whitney statistic for two sample problems, which is equivalent to the Wilcoxon rank statistic (Bamber, 1975). We use ten-fold cross validation (Hastie, Tibshirani and Friedman, 2001) to calculate the AUC measure based on logistic models for each of the *k* clusters to correct for the bias that occurs when the same data is used to test the accuracy of the model and fit the model.

A potential problem that can occur with partitioning data and model building is that exact fits are possible. With logistic regression models this is due to small sample sizes or regions that are almost all extirpated or all present, i.e., the response is uniform over the sub-region. When more than 90 percent of the response is either extirpated or present, instead of fitting logistic model to it, we set this cluster aside and assign a perfect AUC value of one to that cluster. The reasoning behind this is that when such an overwhelming response subregion is discovered, it may indicate an interesting general feature of the area and does not require a separate model (essentially we assign an intercept model to that data). We do include the AUC information into the overall criterion and we define the

final performance measure AUC of the clustering scheme to be the weighted average of the individual AUC measures of each cluster:

$$\text{AUC}=\frac{\sum_i AUC_i n_i + \sum_j n_j}{N}$$

Where *j* indexes the overwhelming clusters and *i* indexes all others. Both *i* and *j* can take values in 0 through *k*. *N* is the total sample size, n_i is sample size for the *i*th regular cluster and n_j is the sample size for the *j*th overwhelming cluster. In addition, we require that the possible clustering solutions will satisfy the requirement that each individual AUC measure will be greater than the benchmark AUC, which is the AUC without any clustering. This guarantees that the final clustering solution will give at least comparable performance as the single model approach. When the condition is not met, it is an indication that clustering is not necessary since a single model will suffice. Therefore, this additional requirement can be viewed as a built-in mechanism to guard against unnecessary clustering.

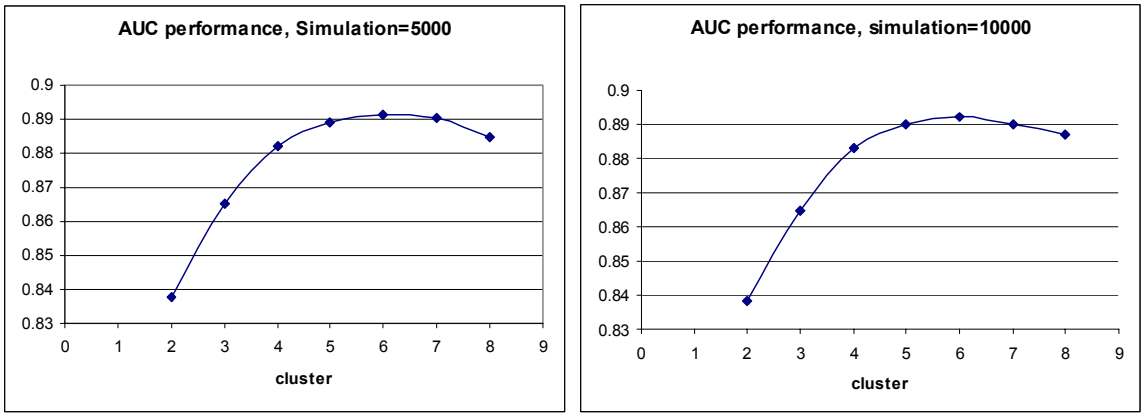
We repeat steps one to three for a range of cluster sizes 2 through a maximum number of clusters, *k,* during one simulation run and repeat the process *S* times, so the total number of simulation runs is *k*S*. We vary the size of *S* to ensure closeness to optimality and evaluate sensitivity of the method.

3.4 Results

3.4.1 Simulation results

Figure 3.2 displays the optimal values of the AUC criterion for cluster 2 through 8 for simulation sizes of 5000 and 10000. The shape of the AUC curve is convex with a maximum at six clusters. The optimal AUC measures for both sets of results are fairly similar (with the difference less than 1 percent) in terms of the AUC criteria and the resulting subregions. A chain of 20000 also produces similar results.

Figure 3.2: AUC performance for 2-8 cluster solutions with 5000 and 10000 simulation runs



3.4.2 Six-cluster solution

3.4.2.1 Geographical layout

We choose the 6 cluster clustering scheme that resulted in max AUC in the 2 simulation sets as the final clustering solution. Figure 3.3 displays the geographical locations of the resulting 6 clusters. The regions are broadly speaking the southern Appalachian mountain region, divided into north and south. West Virginia sites are combined with sites from western Pennsylvania; eastern Pennsylvania, New Jersey and southeastern New York sites; and Western New York sites form a separate region. The final region is defined by sites in eastern New York, Massachusetts, Connecticut and Rhode Island. Cluster 1 was identified indirectly by Thieling (2006) as an area with lower classification rates than other regions.

3.4.2.2 AUC performance

The result of the AUC performance for each cluster is listed in Table 3.2. The overall AUC value is 0.89 which is the weighted average of each individual clusters (0.88 if an AUC value of 0.9, the proportion defining an overwhelming cluster, is assigned to cluster 6). Note that the AUC values are based on ten-fold cross validation results. This large value of AUC indicates good predictive power of our logistic models. To see the improvement over the conventional modeling approach, we did 10-fold cross validation using the logistic regression model with the same 4 predictor variables on the entire dataset, which results in an AUC value of 0.77. Our model-based approach results in a 16% improvement over the conventional single model approach in terms of the predictive capability based on this AUC measure. Another advantage of the model-based approach is its ability to find those "overwhelming" clusters. In our case, cluster 6 turns out to be such a cluster with 90 percent data points being "present". This is a useful discovery from the management perspective as it indicates that brook trout are self-sustaining in almost every subwatershed in that region, and suggests a strategy to maintain and prevent rather than restore.

Figure 3.3: Geographical layout of the optimal 6 clusters

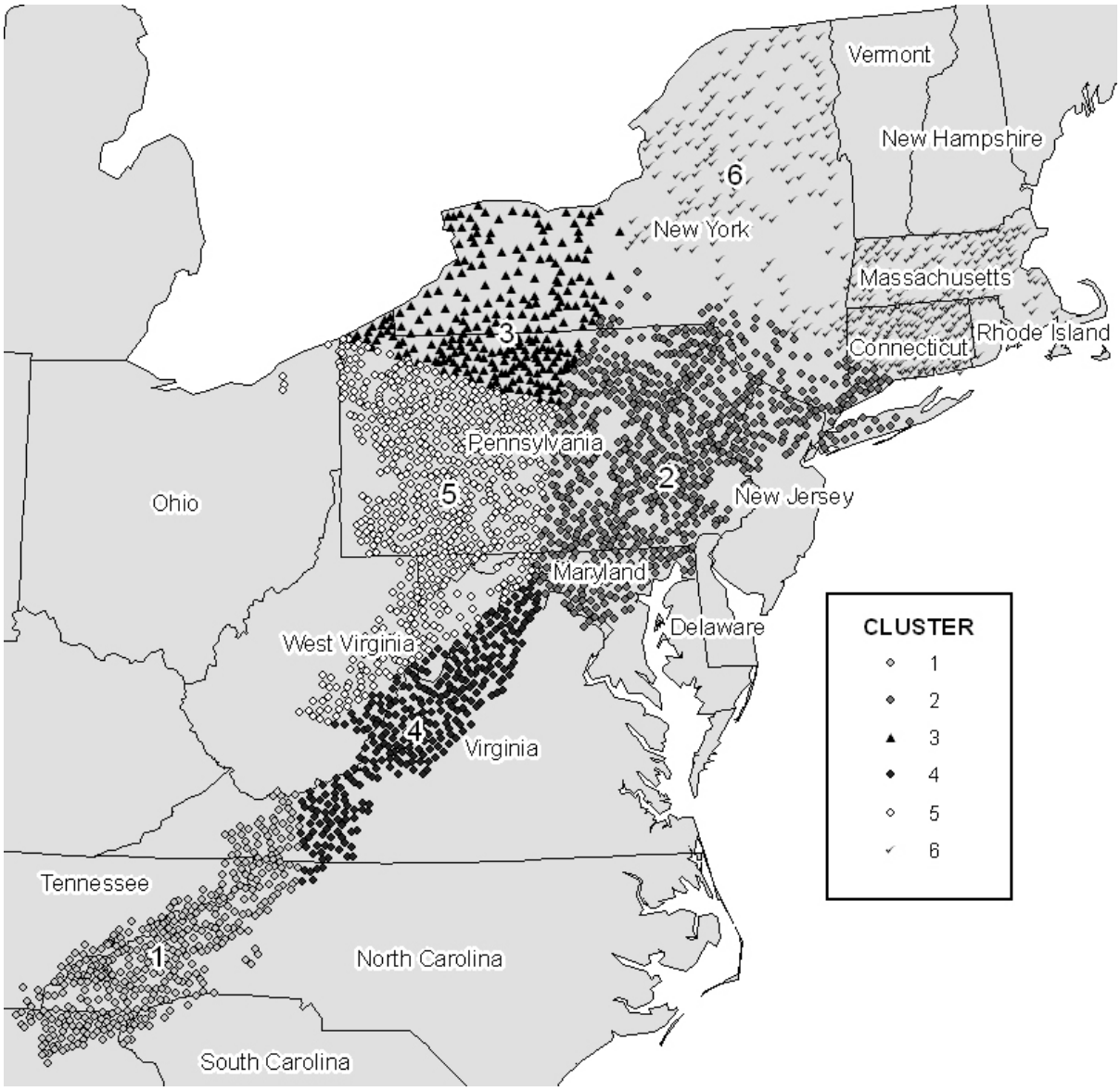


Table 3.2: 10-fold cross validation results for the final 6-cluster solution

Model-based clustering				
Cluster	No.obs	AUC		AUC *
1	403	0.81		0.81
2	757	0.85		0.85
3	251	0.93		0.93
4	314	0.87		0.87
5	633	0.91		0.91
6	431	1.00		0.90 *
Combined	2789	**0.89**		**0.88**

Single model approach	
No.obs	AUC
2789	**0.77**

* AUC when a proportion of 0.90 is used in defining an overwhelming cluster.

3.4.2.3 Parameter estimates

The parameter estimates along with indication of significance of the logistic regression models for the model-based clustering and single modeling approaches are summarized in Table 3.3. We see differences in both magnitudes and signs of those coefficient estimates between the single model and the 6 separate models. There are variations among the 6 clusters as well. In particular, we observe the following: 1) Based on the single model, the parameter estimate for elevation is 0.21 which is relatively small and positive. Elevation is partly an indicator of the water temperatures within the subwatershed. Based on the estimate of a positive number, 0.21, the interpretation is that the colder the subwatershed, the more likely brook trout will be extirpated (recall we are modeling the probability of being extirpated so positive coefficients indicate increasing probability with increasing values of the variable). This is contradictory to the fact that brook trout have a preference for cooler streams. When we look through the estimates obtained from the model-based clustering approach, elevation has a uniformly negative effect on brook trout being extirpated. This is sensible and confirms experts' findings (Hudy, et al., 2006). 2) There are considerable differences in the intercept terms. The

exponential of the coefficient represents the probability of extirpation at sites that have zeros for all variables and are baseline models. While the baseline models (with only intercept term) for most clusters (cluster 2, 3, 5, 6) favor brook trout presence, cluster 1 has a large positive intercept of 2.76. Cluster 1 is the southern Appalachian area that has a high rate of wrongly classifying "extirpated" as "present". A closer look at the original data in that region reveals some interesting features. See Figure 3.4 for the comparisons of the 4 predictor variables between cluster 1 and the remaining clusters. This is an area that has higher elevation (mountain areas), lower agricultural activities and higher percentage of total forested lands compared to other study areas. Based on experts' knowledge, this area will be favorable to brook trout's presence. Therefore, the baseline model should have a negative intercept indicating brook trout abundance (which, we suspect is the cause for the relatively high misclassification rate in that area). In fact, there is no such "presence" dominance in this area. The sample size for "extirpated" and "present" is 201 and 202 respectively. Our model produced a positive intercept estimate of 2.76 for this cluster which implies the general pattern in the area is different from the other areas and further investigation is needed to achieve better classification performance. Based on a retrospective study by Thieling (2006), it turns out that in that area, past land use practice and subsequent stocking of exotic rainbow trout into restored subwatersheds have displaced brook trout and therefore this exotic species is a very important metric affecting brook trout status that, unfortunately, is not among the original 63 metrics compiled. By using model-based clustering, we are able to discern such irregular regions among vast data, and help better control the misclassification rate as compared to a single model approach. In future studies, including an exotic fishes metric in the model may further enhance the model's predictive performance. 3) According to Thieling's (2006) retrospective study, The majority of subwatersheds predicted to be "extirpated" but in fact "present" were located in eastern Pennsylvania and western New York, which correspond to clusters 2 and 3 respectively in our clustering solution. In Thieling's study, high road density along with increasing urbanization resulted in many of the misclassifications. After clustering, the importance of elevation outweighs that of road density. For cluster 2, the estimate (standard error) for elevation and road density are

-1.24 (0.29) and 0.31 (0.15) respectively. For cluster 3, the estimate (standard error) for elevation and road density are -4.56 (1.17) and 1.35 (0.43) respectively. This tells us that in those areas, elevation has more effect on brook trout status than road density. While based on the single model, the importance of elevation is somewhat overlooked resulting in such high misclassification rates.

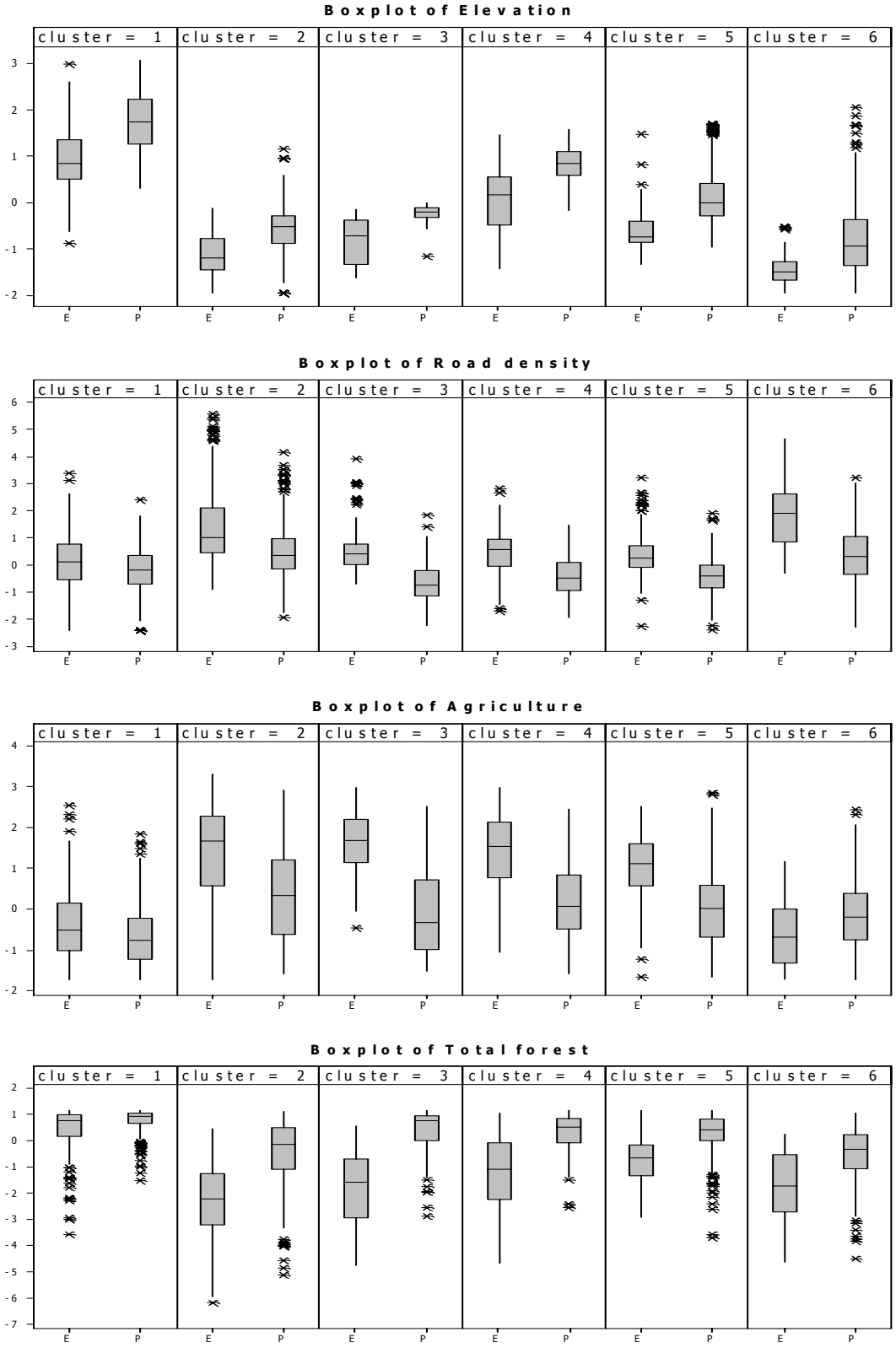
Given the resulting 6 clusters, we further refine the logistic regression models by doing variable selection within each cluster. More details are listed in Appendix B and Appendix C.

Table 3.3: Parameter estimates for logistic models for two modeling approaches

Parameter estimates for logistic model on the entire dataset						
N	# of subwatersheds present/extirpated	Intercept	Elevation	Road density	Agriculture	Total forest
2789	1717/1072	-1.02 **	0.21 **	0.45 **	0.49 **	-0.35 **
Parameter estimates for logistic models based on 6-cluster solution						
cluster	# of subwatersheds present/extirpated	Intercept	Elevation	Road density	Agriculture	Total forest
1	202/201	2.76 **	-1.95 **	0.37 *	-0.5	-0.64
2	403/354	-2.54 **	-1.24 **	0.31 *	0.51 **	-0.42 **
3	150/101	-3.17 **	-4.56 **	1.35 **	1.99 **	1.32 *
4	187/127	0.80 *	-2.62 **	-0.16	0.11	-0.78
5	383/250	-1.91 **	-2.85 **	0.84 **	1.08 **	0.26
6	392/39	-4.84 **	-1.2	0.75 *	-0.16	-0.17

* significance level ≤ 0.05
** significance level ≤ 0.005

Figure 3.4: Box plots of predictor variables for cluster 1 compared to other clusters, using transformed, centered and scaled variables



3.5 Discussion

In this chapter, we propose a method for using model-based clustering on spatial data, in particular, we develop algorithms for segmenting binary response data using logistic regression. Voronoi tessellation techniques are employed and AUC is used as the performance measure during the Monte Carlo search for the optimal clustering solution. Application of this method on a brook trout data set demonstrates the potential of our method for achieving better classification performance than a similar model that ignores clustering.

The models we obtained using model-based clustering can be used in several ways. First, the discovery of the "overwhelming" cluster number 6 suggests that preservation and maintenance may be the correct strategy since brook trout are self-sustaining in almost every subwatershed. Second, irregular pattern of brook trout status is discovered in the Southern Appalachian area in opposition to expert's expectation. Further investigation is needed to check into other potential predictors to better manage that area. Third, well performing predictive models in some study areas can be used to predict future subwatersheds of interest. Fourth, the interpretation of the resulting logistic models can aid in managerial actions combined with other professional knowledge. Our models demonstrate the importance of elevation in all 6 clusters. The large values and the negative signs of the coefficient estimates are an indication of lesser human disturbance at higher elevations. Even though elevation is not a metric that the land managers can control, it suggests that the importance of management of high elevation streams for maintaining self-sustaining brook trout populations. For the other three predictor variables, their differential significance and estimated values in each cluster can aid land managers in setting priorities in making protective management decisions.

We dealt with binary response data in this paper. The method can be extended to situations when the response is multinomial (more than two levels).

Users of our method can easily modify the criteria to achieve various research goals. Recall we use the weighted average AUC as the performance criterion since we hope to obtain better overall classification compared to the single model approach. If one is interested in finding a "hotspot" within the region where the model used can describe the stressor-response relationship nearly perfectly, one can use the maximum AUC as the criterion. We also added additional requirements in the implementation so that each individual AUC value should be greater than the benchmark AUC value. In situations where regions of weak relationships are naturally expected, one can remove that requirement.

We use longitude and latitude as the clustering variables. Other natural environmental variables can also be used as long as they have discriminating power. Bates Prins et al. (2006) use elevation and stream width as the clustering variables in a water quality study in the Mid-Atlantic Highlands. Our work can be easily extended to other application areas as well as long as partitioning the entire dataset makes intuitive sense and there are clustering variables available.

Chapter 4 Hierarchical Models Using Results from Model-Based Clustering

In environmental monitoring, data are sometimes collected over large spatial regions. While all the data are collected to address the same problem, inevitable differences exist among different study regions. Therefore, a single model assuming homogeneity across the different study regions may not be appropriate. Possible alternative approaches include separate modeling approach and hierarchical modeling approach. To apply hierarchical models, it is important to determine the hierarchical structure of the data. Sometimes, the hierarchical or clustered structure is obvious. For example, animal and human studies of inheritance deal with a natural hierarchy where offspring are grouped within families (Goldstein, 1995). In some environmental monitoring studies where continuous spatial regions are involved, the clustered structure may not be apparent, as in the case of the brook trout study. In Chapter 2, we developed a hierarchical model using an empirical hierarchical structure constructed by grouping data points in several neighboring states into the same cluster. In Chapter 3, we developed a model-based clustering method and obtained a more reasonable clustering structure of the data. In this chapter, we integrate the clustering results obtained in Chapter 3 to develop an improved Bayesian hierarchical model. The classification results produced by such a model are better than those produced by other hierarchical models with empirical clustering structures.

4. 1 Hierarchical models vs. non-hierarchical models

Many observational and experimental data used in environmental studies, social sciences, medical research, human and biological sciences have a hierarchical or clustered structure. When a one-level single model is used to model the entire data as if they come from the same underlying distribution, possible differences among clusters are overlooked and may result in misleading results. A well-known example is a study of elementary school children carried out in the 1970's. The study by Bennett (1976) claimed that children exposed to the so-called 'formal' styles of teaching showed more progress than those who were not. The data were analyzed using traditional multiple regression techniques where a one-level single regression model was used and the grouping of children within teachers/classes were ignored. The results were statistically significant. Later, Aitkin et al, (1981) demonstrated that when the analysis properly accounted for the grouping of children into classes, the significant differences of the "formal" teaching styles disappeared. This example illustrated the importance of hierarchical modeling when a hierarchical structure of the data is reasonably suspected.

4.2 Hierarchical structure of data

As argued earlier, when data display a hierarchical structure, it is reasonable to apply hierarchical models. A question that naturally arises is how to determine the actual hierarchical or clustered structure of the data.

The literature on hierarchical models gives very little suggestion on this matter. They generally assume that the hierarchical structure is given, as Goldstein (1995) pointed out that they are concerned "only with the *fact* of such hierarchies not their provenance".

This assumption is quite reasonable in cases where the hierarchical structure is obvious, as in the example of animal and human studies of inheritance where families are the natural clusters. There are, however, situations when the hierarchical structure is less than

apparent. For example, in environmental water quality monitoring, data may be collected in a vast spatial region covering many states. It is reasonably suspected that the homogeneity assumption of the entire data will hardly hold. Therefore, it is natural to model the data hierarchically. There are several options to determine the actual hierarchical structure to use. One approach is to make use of information such as the relative geographical location of the data points (south, north, east or west), the established ecoregions each data point belongs to, or information on natural environmental factors with differentiating power (elevation, temperature, etc) to help determine the empirical hierarchical structure of the data. The resulting hierarchical structure may not be optimal, but it should be fairly reasonable if done properly. Or one can develop a more formal procedure to investigate the clustering structure of the data so that a more sophisticated hierarchical structure can be obtained based on certain criteria.

Because the brook trout study involves a vast spatial region and a large number of states, we propose a hierarchical modeling approach to properly account for the similarities and differences in different study regions. We developed a model-based clustering method in Chapter 3 to obtain a more sophisticated clustering structure so that the relationship between variables are similar within the same cluster and dissimilar among different clusters. In this chapter, we develop a hierarchical model using this more sophisticated clustering structure and compare the resulting classification results to those produced by two other hierarchical models using empirical clustering structures.

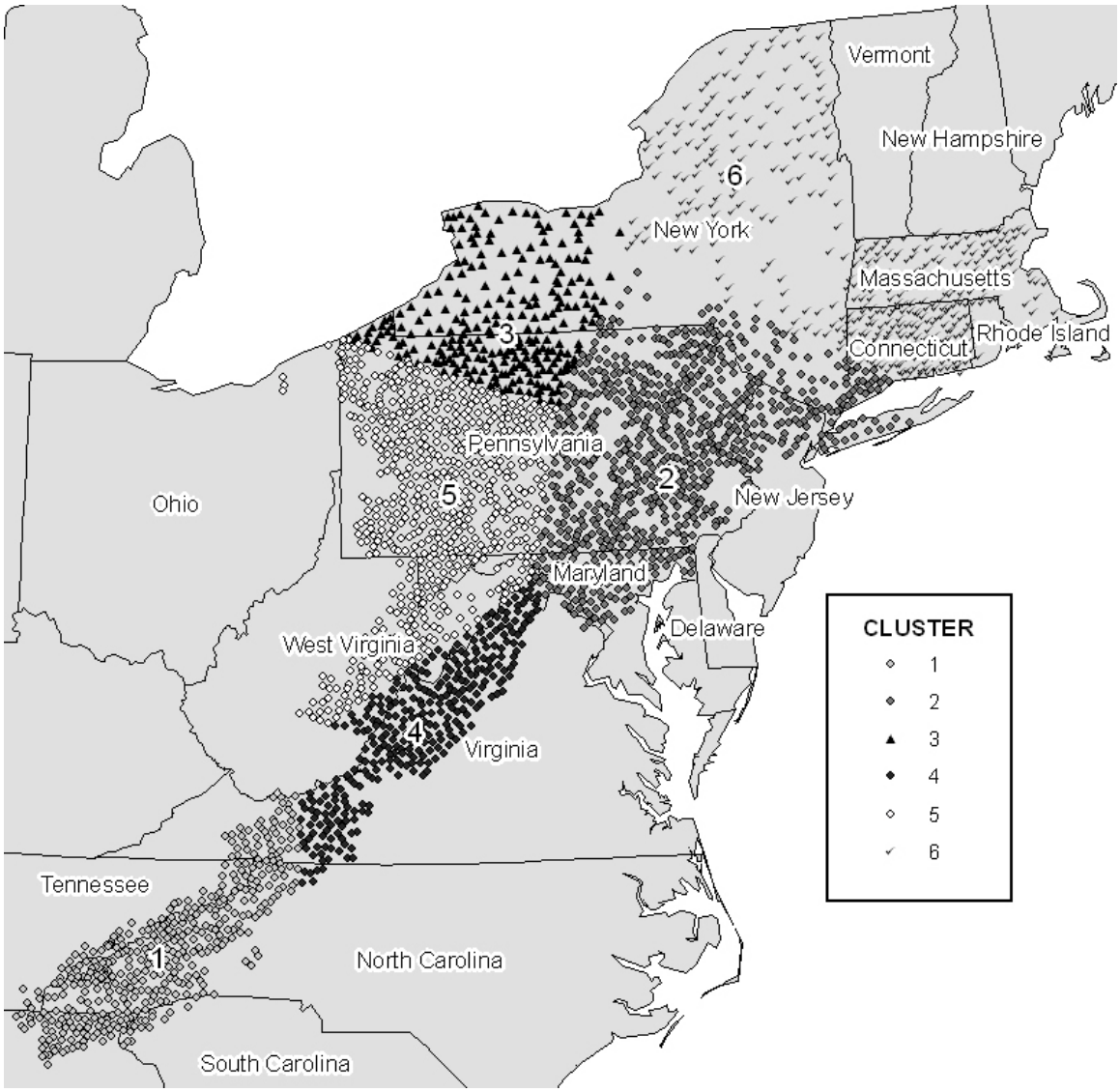
4.3 Hierarchical model with improved hierarchical structure

4.3.1 Model-based clustering structure

In Chapter 3, the six clusters are formed based on the relationship between the response and the environmental variables and are therefore more reasonable. Due to the complete separation problem encountered in region 1, only 2789 data points in the other 4 regions are used in Chapter 3 instead of the original sample size of 3337. Since we decide to use the clustering results in Chapter 3 to develop the hierarchical model here, the same data

set used in the model-based clustering is used with the same 4 predictor variables to build the classification model. The graphical layout of the 6 clusters is illustrated in Figure 4.1 below.

Figure 4.1: Hierarchical structure of the brook trout data based on clustering result. (*N*=2789)



4.3.2 Hierarchical model

4.3.2.1 Model

The setup of the hierarchical model is the same as in Chapter 2 except that the number of dimensions of the parameter vectors changed from six to five and the number of clusters changed from five to six. Below is the model.

$$
\begin{aligned}
&Y_{ij} \mid p_{ij} \sim Bernoulli\,(p_{ij}) \\
&\text{logit}(p_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}_i \\
&\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \Sigma \sim N_k(\boldsymbol{\mu}, \Sigma^{-1}) \\
&\pi(\boldsymbol{\mu}) \sim N_k(\boldsymbol{\alpha}, \Omega) \\
&\pi(\Sigma) \sim Wishart(\mathrm{R}, v)
\end{aligned}
$$

where $i = 1,2,...,6$ indexes clusters and $j = 1,2,...,n_i$ indexes observations of a cluster. The *j*th observation in the *i*th cluster Y_{ij} is assumed to follow a Bernoulli distribution with event probability p_{ij}. A logistic regression model is used to model p_{ij} with the link being the logit function and the linear predictor $\mathbf{x}_{ij}'\boldsymbol{\beta}_i$. The $\boldsymbol{\beta}_i$'s are the parameter vector for each cluster (we have six clusters). The dimension of $\boldsymbol{\beta}_i$ is five (we have four predictor variables plus an intercept term in the model). The $\boldsymbol{\beta}_i$'s are assumed to come from the same multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and precision matrix $\sum$. The hyper prior distribution for $\boldsymbol{\mu}$ is the commonly used multivariate normal distribution. The hyper prior distribution for $\sum$ is Wishart distribution. For more details, please refer to Chapter 2.

4.3.2.2 MCMC simulation

Due to the complexity of the model, the posterior distributions of the parameters of interest are not available in closed form. The Markov chain Monte Carlo (MCMC)

method is used to implement this logistic hierarchical model. In particular, a hybrid of the Gibbs sampling (Gelfand and Smith, 1990) and the Metropolis algorithm (Metropolis, 1949, 1952) is used to obtain the posterior distributions of the parameters of interest. The algorithm is similar to what was used in Chapter 2 and the details are not repeated here.

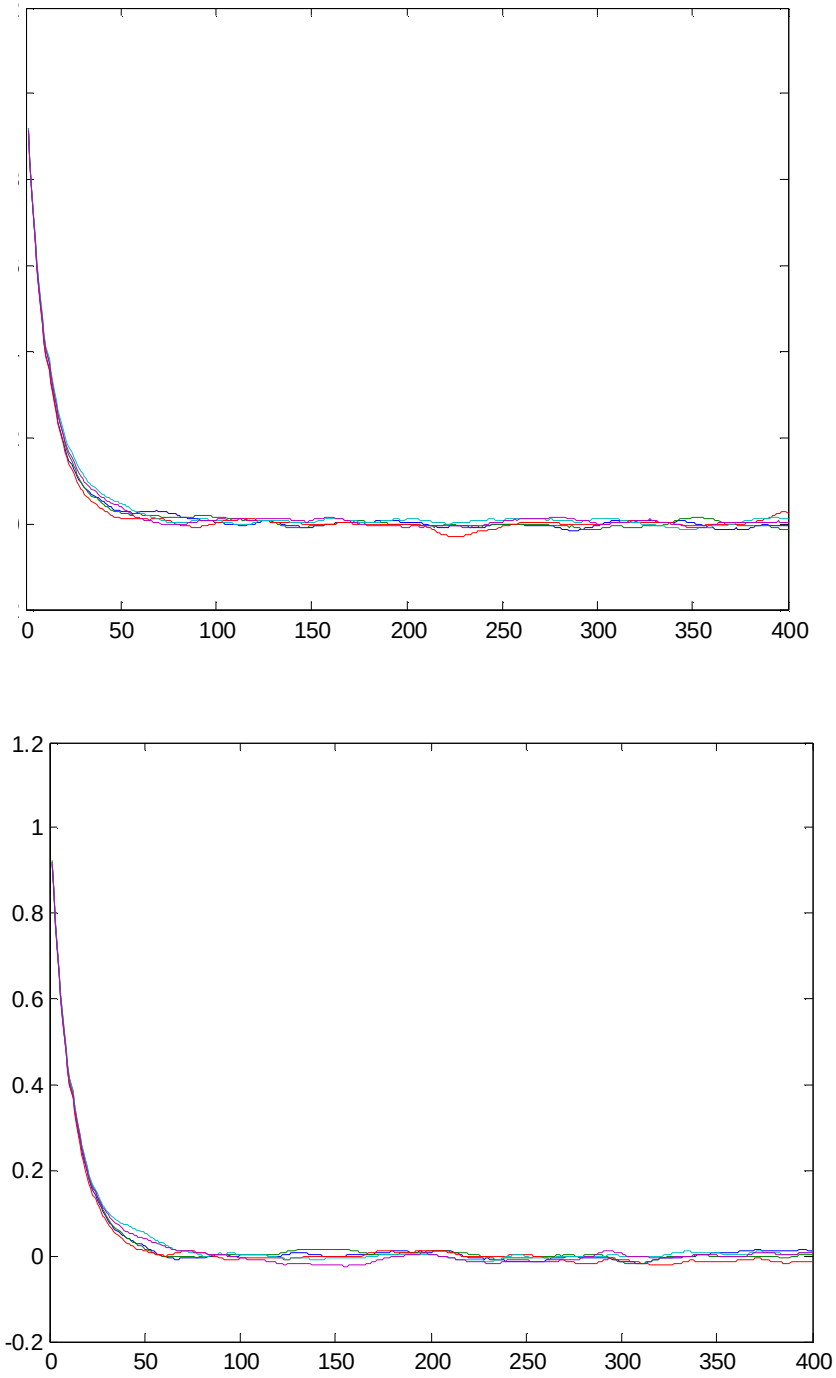
The simulation is implemented in MATLAB. Two chains of simulations are used with each run of size 155,000. For each chain, a burn-in of 5000 runs is used to eliminate the effect of starting values and a thinning of 50 is used to ensure that the auto-correlation drops to nearly zero. Graphical diagnostics including auto-correlation plots and trace plots (see Figures 4.2, 4.3 and 4.4 for demonstration) show the simulation process has converged. Two formal diagnostic tools, the Potential Scale Reduction Factor (PSRF) (Gelman and Rubin, 1992) and the Multivariate Potential Scale Reduction Factor (MPSRF) (Brooks and Gelman, 1998), also suggest convergence. The MPSRF values for $\boldsymbol{\beta}_i$'s, $\boldsymbol{\mu}$ and $\sum$ are listed in Table 4.1. The final analysis is based on 6000 samples combined from two chains after burn-in and thinning.

Table 4.1: Multivariate Potential Scale Reduction Factor

MPSRF	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$	$\boldsymbol{\beta}_5$	$\boldsymbol{\mu}$	Σ
Parameter	1.0001	1.0004	1.0005	1.0005	1.0002	1.0015	1.0015

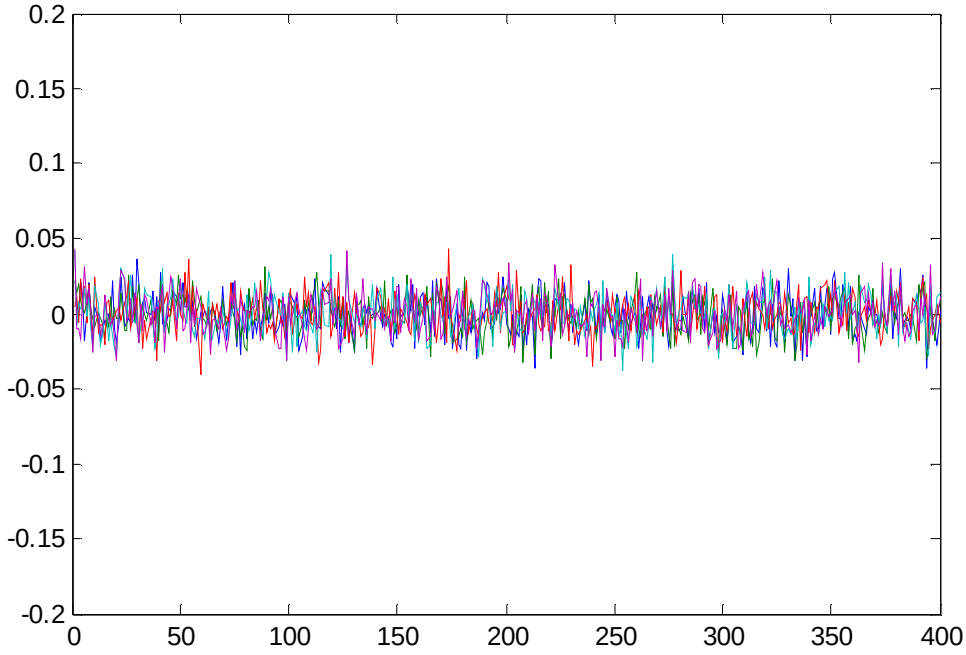
Note: Values below 1.1 indicate convergence.

Figure 4.2: Auto-correlation plots of parameters for cluster 1 from chain 1 and chain 2



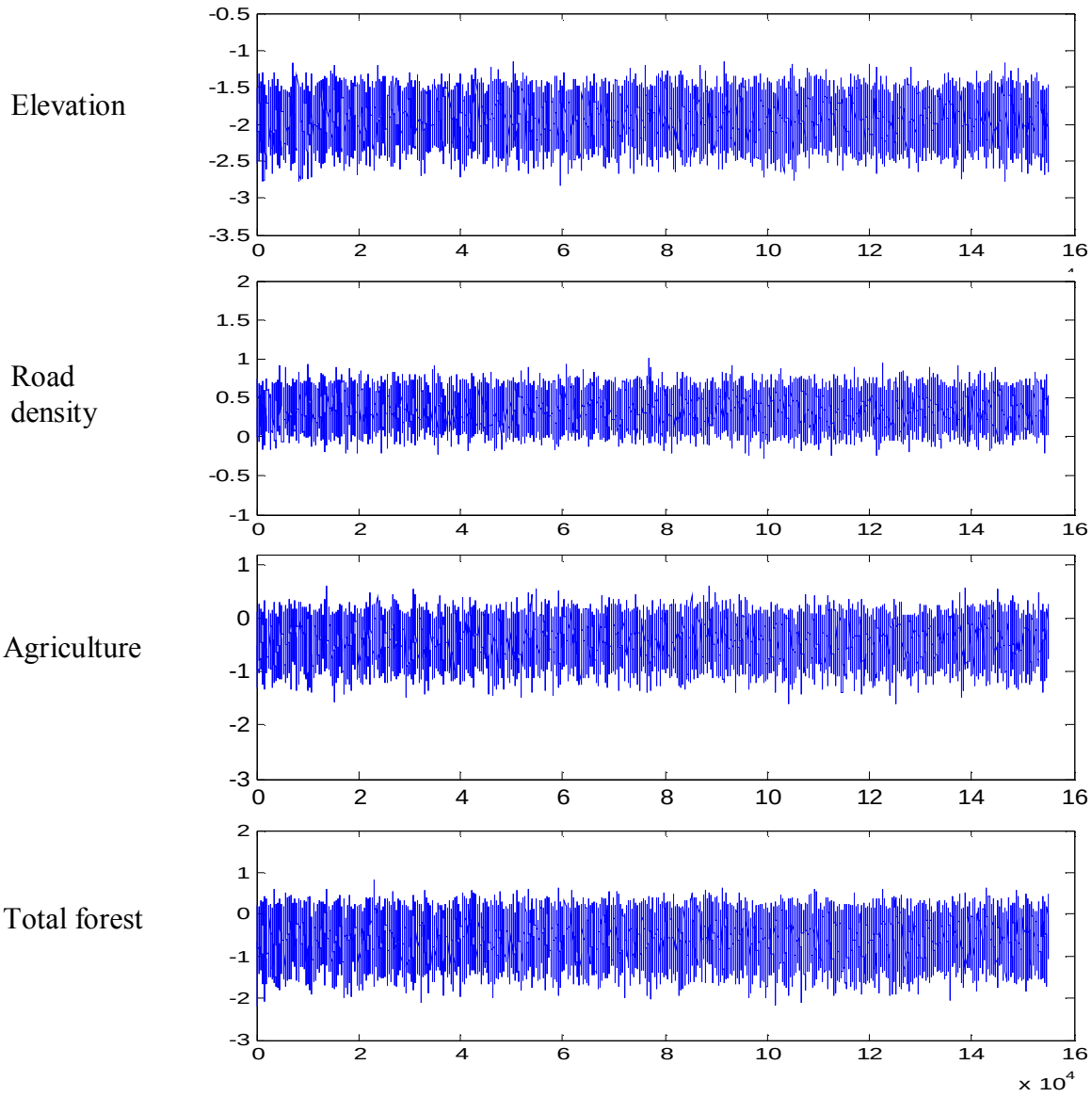
Note: The plot on the top is from chain 1 and the one on the bottom is from chain 2. Plots for other clusters are similar and are omitted. Note the auto-correlation drops to nearly zero after lag 50.

Figure 4.3: Auto-correlation plot of parameters for cluster 1 after a burn-in of 5000 and thinning of 50



Note: the auto-correlation drops to nearly zero.

Figure 4.4: Trace plots of the coefficients for the 4 predictor variables for cluster 1



Note: The variable name is indicated on the left side of each plot. The y-axis is the simulated value of the coefficient and the x-axis is the iteration number. The rest of the plots show similar patterns and are therefore omitted.

4.3.3 Results

4.3.3.1Parameter estimates

We obtain estimates (specifically, posterior means) for each $\boldsymbol{\beta}_i$ using the Bayesian hierarchical logistic modeling approach (See Table 4.2 for the results). We also apply a one-level single logistic model on the entire data to obtain estimates (See Table 4.3 for the maximum likelihood estimates).

Table 4.2: Regression coefficient estimates for the 6 clusters from the Bayesian hierarchical logistic modeling approach

	Region					
	1	2	3	4	5	6
Var	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
Int	2.71 (0.33)	-2.61 (0.24)	-2.26 (0.42)	0.57 (0.32)	-1.77 (0.22)	-4.71 (0.65)
Elev	-1.93(0.22)	-1.35 (0.27)	-2.62 (0.66)	-2.34 (0.39)	-2.65 (0.28)	-1.13 (0.49)
Road	0.35 (0.16)	0.32 (0.14)	0.99 (0.33)	0.06 (0.23)	0.80 (0.19)	0.80 (0.32)
Ag	-0.41 (0.28)	0.52 (0.12)	1.12 (0.41)	0.27 (0.37)	0.90 (0.29)	-0.03 (0.28)
Forest	-0.57 (0.39)	-0.40 (0.12)	0.20 (0.43)	-0.54 (0.37)	0.04 (0.30)	-0.15 (0.26)

Note: mean-------posterior mean sd--------standard deviation
Var---------Variable Int------- Intercept
Elev--------Elevation Road-----Road density
Ag---------Agriculture Forest----Total forest

Table 4.3: Maximum likelihood estimates of regression coefficients from the one-level single logistic model using all the data

All regions	
Variable	**Estimate (s.e)**
Intercept	-1.02 (0.05)
Elevation	0.21 (0.06)
Road density	0.45 (0.06)
Agriculture	0.49 (0.06)
Total forest	-0.35 (0.07)

Note: s.e------standard error

Results in Table 4.2 suggest that the signs, magnitude and significance of the coefficient estimates vary across the 6 clusters in the hierarchical modeling approach. Overall, each region shows certain degree of dissimilarities.

All four predictor variables are highly significant using the one-level single logistic model. Using hierarchical modeling, the relative importance of each predictor becomes clearer in each cluster, which can help land managers in setting up different management priorities in different areas.

Results concerning elevation in this improved hierarchical model are similar to those in the empirical hierarchical model in Chapter 2. From the one-level single logistic model, the coefficient estimate for elevation is positive 0.214. Elevation is partially an indicator of the temperatures in the watersheds. The positive estimate of elevation suggests that the colder the place is (with higher elevation), the more likely brook trout will be extirpated. This is contradictory to the fact that brook trout have a preference for cooler streams.

When we look at the estimates obtained from the hierarchical modeling approach, elevation has a uniformly negative effect on brook trout being extirpated.

4.3.3.2 Classification performance

We are most interested in the model's predictive ability. Two classification performance measures are computed for both the hierarchical model and the one-level single model. They are AFCCF (Average Fraction of Correctly Classified for Fit) and AUC (the area under the ROC curve). The details of these two measures are given in previous chapters and are not repeated here.

In the Bayesian hierarchical model, the posterior distribution of $\boldsymbol{\beta}$ is obtained. The posterior mean and/or posterior median are commonly used as the point estimate for $\boldsymbol{\beta}$. For the hierarchical model, the predicted probabilities are calculated using $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{posterior mean}})}$. In the one-level single model, $\hat{y} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}$ is used as the predicted probability where $\hat{\boldsymbol{\beta}}$ is the MLE.

The AFCCF and the AUC results for the two modeling approaches are summarized in Table 4.4.

Table 4.4: AFCCF and AUC measures for the hierarchical model and the one-level single model

Modeling approach	AFCCF	AUC
Bayesian hierarchical model	0.75	0.90
One-level single model	0.63	0.77

Note: *N*=2789. Larger values indicate better classification accuracy.

Compared with the one-level single model, the hierarchical model improves the classification accuracy by 18% in terms of the AFCCF measure and by 16% in terms of the AUC measure.

4.3.4 Comparison with the single model with dummy variables

We also fit a single model using dummy variables indicating the cluster membership of each data point and calculate the predicted probabilities. That is, we allow different intercept terms for each cluster in the model but the coefficients for the predictor variables are the same for all the clusters. The advantage of this approach over the one-level single model approach is that the baseline differences among the clusters are taken into account. The resulting coefficient estimates are listed in table 4.5.

Table 4.5: Maximum likelihood estimates of regression coefficients from the single logistic model with dummy variables

Parameter	Estimate	s.e
Intercept	-1.34	0.07
Cluster 1	4.48	0.23
Cluster 2	-1.74	0.14
Cluster 3	-0.40	0.16
Cluster 4	1.66	0.17
Cluster 5	0.03	0.11
Elevation	-2.04	0.14
Road density	0.33	0.08
Agriculture	0.43	0.08
Total forest	-0.35	0.09

The classification performance for both hierarchical model and the single model with dummy variables are summarized in Table 4.6. The classification performances for the two modeling approaches are quite comparable.

Table 4.6: AFCCF and AUC measures for the hierarchical model and the single model with dummy variables

Modeling approach	AFCCF	AUC
Bayesian hierarchical model	0.75	0.90
Single model with dummy variables	0.73	0.89

Note: *N*=2789. Larger values indicate better classification accuracy.

4.3.5 Comparison with other empirical hierarchical models

With no other information available, we used an empirical hierarchical structure in the hierarchical model in Chapter 2. The model did improve classification performance over the non-hierarchical model. To do a fair comparison of the two hierarchical models' classification performance, we modified the hierarchical model in Chapter 2 by using only the 4 regions as the clusters so that the sample size is the same for both hierarchical models.

Another empirical clustering structure can be obtained by simply using the original 13 states as the clustering units. In spatial studies, this may be the simplest approach since this information is readily available. We combine the observations in the state Ohio (only 4 of them) with Pennsylvania and develop a Bayesian hierarchical model using the resulting 12 states as the clusters.

A potential problem with the grouping schemes used in these two empirical hierarchical structures is that the states are used as the grouping units which by definition are administrative districts not environmental units. For studies involving rivers and streams as in the brook trout study, the watersheds close to the same river or stream should naturally be in the same cluster. If the river runs through several states, then those related watershed could be assigned to different clusters.

The modeling and implementation steps of these two hierarchical models are similar to those illustrated earlier in this chapter, and are therefore omitted.

We calculate the AFCCF and AUC measures for these two empirical hierarchical models. We summarize all three hierarchical models' classification performance improvements compared to the single model approach in Table 4.7.

Table 4.7: Classification performance comparison of three hierarchical models

Modeling Approach		AFCCF	AUC
Single model		0.63	0.77
Hierarchical model clustering unit: 12 states		0.69	0.85
	Improvement *	**0.09**	**0.10**
Hierarchical model clustering unit: 4 regions		0.70	0.86
	Improvement *	**0.11**	**0.10**
Hierarchical model clustering unit: 6 clusters		0.75	0.90
	Improvement *	**0.18**	**0.16**

Note: the improvement is based on the comparison with the single model approach. *N*=2789.

As Table 4.7 suggests, the hierarchical model using the 12 states as the clusters improved the model's AFCCF performance by 9% and improved the AUC performance by 10% compared to the single model approach. The improvements of the hierarchical model with the empirical 4 regions as the clusters are 11% for the AFCCF measure and 10% for the AUC measure, whereas the improvement of the hierarchical model with the model-based clustering structure for these two measures are 18% and 16% respectively. The hierarchical model with the model-based clustering structure had the greatest improvement among the three hierarchical models which shows that there is benefit to selecting clusters when feasible.

4.3.6 Concluding remarks

In this chapter, we utilized the clustering structure obtained in Chapter 3 to improve the hierarchical model we developed in Chapter 2. This improved hierarchical model showed highly significant improvements in terms of classification performance over the non-hierarchical model.

Compared to two other hierarchical models using empirical hierarchical structures, the hierarchical model with a more sophisticated hierarchical structure used in this chapter improves classification ability to a greater extent. The results support our argument that when a hierarchical structure of data is reasonably suspected but the actual hierarchical structure is unknown, discovering a reasonable data structure before applying the hierarchical model could be beneficial.

Despite their usefulness, hierarchical models need to be used with care. In some circumstances when there is little structural complexity, the one-level single model may suffice. This agrees with our point earlier that some research on the data structure is recommended before applying the hierarchical model.

Chapter 5 Future research

The model-based clustering method developed in Chapter 3 deals with binary response data. The method can be extended to situations when the response is multinomial (more than two levels). In practice, multinomial response data are often encountered. In the brook trout study, originally the response has three levels: extirpated, reduced and intact. Clustering models with the ability to differentiate between all three response levels are needed. A potential problem is that several different models may be required (e.g., generalized logit model versus cumulative logit model).

We used AUC as the criterion in developing the model-based clustering method in classification analysis. A future research area could involve adapting partial AUC instead of the usual AUC as the performance criterion. The usual AUC measure summarizes the model's discriminant ability across all cutoff values, including some cutoffs that are rarely used in practice. For example, the rightmost area under the curve corresponds to very high (nearly 1) false positive rate and very high true positive rate since it virtually classify all observations to be present (event). The same is true with the lower left corner of the ROC curve where virtually all observations are classified as absent (non-event). The model's performance in those ranges of cutoff values is of little practical importance (Dodd and Pepe, 2003).

In model-based clustering, the same set of predictors is used in building the logistic models for all clusters. We argue that a modified approach that incorporates a variable selection step into the clustering process could be more flexible and offer richer models to the users. We need to be very careful, though, in dealing with the potential

computational difficulty of complete separation that may occur during the variable selection step.

The Bayesian hierarchical modeling approach achieves great success in terms of boosting the classification performance compared to the one-level single modeling approach. It, however, improves very little over the single model approach when the clustering information is considered. More work is needed to establish conditions under which the hierarchical model is needed and would result in more noticeable gain over the other modeling approaches.

References

Agresti, A. (2002), *Categorical Data Analysis*, second edition, John Wiley and Sons, Inc.

Albert, A. and Anderson, J.A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1-10

Aitkin, M., Anderson, D. and Hinde, J. (1981), Statistical Modelling of Data on Teaching Styles (with discussion), Journal of the Royal Statistical Society. Series A., 144, 148-161.

Bamber, D. (1975), "The Area above the Ordinal Dominance Graph and the Area below the Receiving Operating Characteristic Graph," *Journal of Mathematical Psychology*, 12, 387-415

Bedrick, E. J., Christensen, R., and Johnson, W. (1997), "Bayesian Binomial Regression: Predicting Survival at a Trauma Center," *The American Statistician*, 51, 211-218.

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," *Journal of the American Statistical Association*, 88, 1149-1159.

Bennett, N. (1976), *Teaching Styles and Pupil Progress*, London, Open Books.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer-Verlag, New York.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press

Birch, J.J. (2002), Stat 5514 class note, department of statistics, Virginia Polytechnic institute and state university

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, John Wiley and Sons, Inc.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and*

Regression trees. Belmont, CA: Wadsworth International Group.

Brooks, S.P. and Gelman, A. (1998), "General Methods for Monitoring Convergence of Iterative Simulations," Journal of Computational and Graphical Statistics. 7, 434-455.

Browne, W. J., and Draper, D. (2004), "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models," Submitted.

Carlin, B. P., and Louis, T. A. (2000), "Empirical Bayes: Past, Present and Future," *Journal of the American Statistical Association*, 95, 1286-1289.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, second edition, Chapman & Hall/CRC.

Chipman, H., George, E. I. and McCulloch, R. E. (2002), "Bayesian Treed Model," *Machine Learning*, 48, 299-320.

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

Collett, D. (1991), *Modelling Binary Data*, Chapman & Hall.

Congdon, P. (2001), *Bayesian Statistical Modelling*, John Wiley and Sons, Inc.

Congdon, P. (2003), *Applied Bayesian Modelling*, John Wiley and Sons, Inc.

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883-904.

Daniels, M. J., and Gatsonis, C. (1999), "Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization," *Journal of the American Statistical Association*, 94, 29-42.

Datta, G. S., and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Application to Small Area Estimation," *Annals of Statistics*, 19, 1748-1770.

Denison, D. G. T. and Holmes, C. C. (2001), "Bayesian Partitioning for Estimating Disease Risk," *Biometrics*, 57, 143-149

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, John Wiley & Sons, LTD.

Dodd, L. E. and Pepe, M. S. (2003), "Partial AUC Estimation and Regression," *Biometrics*, 59, 614-623

Efron, B. (1996), "Empirical Bayes Methods for Combining Likelihoods," *Journal of the American Statistical Association*, 91, 538-550.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and its Applications*, London:

Chapman and Hall

Fay, R. E., and Herriot, R. A. (1979), "Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Friedman, J. H (1991), "Multivariate Adaptive Regression Splines" (with discussion). *Annals of Statistics* 19, 1

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulations Using Multiple Sequences," *Statistical Science*, 7, 457-472.

Gelman, A., and Little, T. C. (1997), "Poststratification into Many Categories Using Hierarchical Logistic Regression," *Survey Methodology*, 23, 127-135.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman & Hall/CRC.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Ghosh, M. and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal," *Statistical Science*, 9. 55-93.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), eds. *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Goldstein H. (1995), *Multilevel Statistical Models*, New York: Halsted Press.

Good, I. J. (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Mehods*, Cambridge, Massachusetts: M.I.T. Press.

Greco, F. P., Lawson, A. B., Cocchi, D and Temples, T. (2005), "Some Interpolation Estimators in Environmental Risk Assessment for Spatially Misaligned Health Data," *Environmental and Ecological Statistics*, 12, 379-395

Green, D. M. and Swets, J.A. (1966), *Signal Detection Theroy and Psychophisics*. New York: John Wiley and Sons, 45-49

Hanley, J.A. and McNeil, B.J. (1982), "The Meaning and Use of the Area under a Receiving Operating Characteristic (ROC) Curve," *Radiology*, 143, 29-36

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Hastings, W. K. (1970), "Monte Carlo Sampling Method Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.

He, Z, and Sun, D. (2000), "Hierarchical Bayes Estimation of Hunting Success Rates with Spatial Correlations," *Biometrics*, 56, 360-367.

He, Z, and Sun, D. (1998), "Hierarchical Bayes Estimation of Hunting Success Rates", *Environmental and Ecological Statistics*, 5, 223-236.

Holmes, C. C., Denison, D. G. T., and Mallick, B. K. (1999), "Bayesian Partitioning for Classification and Regression," Technical Report, Imperial College, London.

Hosmer, D. W., and Lemeshow, S. (2000), *Applied Logistic Regression*, John Wiley & Sons, INC.

Hudy, M., Thieling, T. M., Gillespie, N. and Smith, E. P. (2006) "Distribution, Status and Perturbations to Brook Trout within the Eastern United States," Final report to the Eastern Brook Trout Joint Venture

Kahn, M. J., and Raftery, A. E. (1996), "Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression with Unobserved Heterogeneity," *Journal of the American Statistical Association*, 91, 29-41.

Kass, R. E., and Steffey, D. (1989), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models," *Journal of the American Statistical Association*, 84, 717-726.

Kass, R. E., and Wasserman, L. (1996), "The Selection of Prior Distribution by Formal Rules," *Journal of the American Statistical Association*, 91, 1343-1370.

Kim, H. M, Mallick, B. K. and Holmes, C. C. (2005), "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," *Journal of the American Statistical Association*, 100, 653-668

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

Li, K. C. and Duan, N. (1989), "Regression Analysis under Link Violation," *Annals of Statistics*, 17, 1009-1052

Lindey, D. V., and Smith, A. F. M. (1972), "Bayesian Estimate for The Linear Model,"

Journal of the Royal Statistical Society, B 34, 1-41.

Ma, S., Huang, J. (2005), "Regularized ROC method for Disease Classification and Biomarker Selection With Microarray Data," *Bioinformatics*, Vol 21, No.24, 4356-4362

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

Meng, C. Y. K., and Dempster, A. P. (1987), "A Bayesian Approach to the Multiplicity Problem for Significance Testing with Binomial Data," *Biometrics*, 43, 301-311.

Metropolis, N., and Ulam, S. (1949), "The Monte Carlo Method," *Journal of the American Statistical Association*, 44, 335-341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M,N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1092

Møller, J. (1994), *Lectures on random Voronoi tessellations. Lecture notes in statistics*, 87, Springer-Verlag

Morris, C. N. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47-55.

Myers, R. H. (1990), *Classical and Modern Regression with Applications* (2^{nd} edition), PWS-KENT Publishing Company

Nichols, S., Sloane. P., Coysh. J., Williams. C., and Norris, R. (2000), "Australian Capital Territory, AUStralian RIVer Assessment System," Cooperative Research Center for Freshwater Ecology, University of Canberra ACT2601

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N (2000), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, second edition, John Wiley & Sons, LTD.

Pepe, M. S. (2000), "Receiving Operating Characteristic Methodology," *Journal of the American Statistical Association*, Vol 95, No.449, 308-311

Pepe, M. S. (2003*), The Statistical Evaluation of the Medical Tests for Classification and Prediction*, Oxford University Press, UK

Pepe, M. S., Cai, T. and Longton, G. (2005), "Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve," Biometrics 0 (0),

-.doi: 10.1111/j.1541-0420.2005.00420.x

Bates Prins, S.C., Smith, E.P., Angermeier, P.L. and Yagow, E.R. (2006), "Clustering Using Stressor-Response Relationships with Discussion on Optimal Criteria," submitted to the *Journal of Computational and Graphical Statistics*.

Raftery, A. E. and Lewis, S. M. (1992), "How Many Iterations of the Gibbs Sampler," *Bayesian Statistics*, 4, 641-649.

Richardson, S., and Gilks, W.R. (1993), "A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models," *American Journal of Epidemiology*, 138, 430-442.

Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons, New York.

Ripley, B. D. (1981), *Spatial Statistics*, Wiley, New York

Robert, G. O. (1995), "Markov Chain Concepts Related to Sampling Algorithms," in *Markov Chain Monte Carlo in Practice*, Chapman & Hall London.

Rogers, C. A. (1964), *Packing and Covering*, Cambridge University press, London.

Scott, S. L., and Ip, E. H. (2002), "Empirical Bayes and Item-Clustering Effects in a Latent Variable Hierarchical Model: A Case Study from the National Assessment of Educational Progress," *Journal of the American Statistical Association*, 97, 409-419.

SYSTAT^R 9, Statistics I, Chapter 17 "Logistic Regression".

Thieling, T. M. (2006), "Assessment and Predictive Model for Brook trout (Salvelinus Fontinalis) Population Status in the Eastern United States", Master thesis, Dept of Biology, James Madison University

Wolpert, R. L., and Warren-Hicks, W. J. (1992), "Bayesian Hierarchical Logistic Models for Combining Field and Laboratory Survival Data," *Bayesian Statistics*, 4, 525-546.

Wong, G. Y., and Mason, W. M. (1985), "The Hierarchical Logistic Regression Models for Multilevel Analysis," *Journal of the American Statistical Association*, 80, 513-524.

Wright, J. F., D. Moss, P. D. Armitage, and M. T. Furse. (1984), "A Preliminary Classification of Running-Water Sites in Great Britain Based on Macroinvertebrate Species and the Prediction of Community Type Using Environmental Data," *Freshwater Biology* 14, 221-256

Appendix A. Key MATLAB codes for Gibbs-Metropolis simulation

% Main function of the Bayesian hierarchical logistic regression model.

```
clear all;
global Data;

% Starting value of Beta vectors for chain1;
chain1_beta=[…];
chain2_beta=[…];

% Stack two chains into a multidimensional  array for Beta vectors;
chain_beta=cat(3, chain1_beta,chain2_beta);

% Starting value of mu vector for chain1;
chain1_mu=[…];
chain2_mu=[…];

% Stack two chains into a multidimensional  array for mu vector;
chain_mu=cat(3,chain1_mu, chain2_mu);


% Starting value of invsigma matrix vector for chain1;
chain1_invsigma=[…];
chain2_invsigma=[…]

% Stack two chains into a multidimensional  array for invsigma matrix;
chain_invsigma=cat(3,chain1_invsigma, chain2_invsigma);

N =105000;   % number of iterations;
% Initialization
beta=[];mu=[];invsigma=[]; accept=[0 0 0 0 0];

for i=1:size(chain_beta,3)              %for the ith chain;
  param_beta= chain_beta(:,:,i);     %use the corresponding starting values for that chain;
  param_mu=chain_mu(:,:,i);
  param_invsigma=chain_invsigma(:,:,i);
  [output_beta,output_mu,output_invsigma,output_acpt]=main_function_estimate(N,
param_beta,param_mu,param_invsigma);
beta=[beta;output_beta];mu=[mu;output_mu];invsigma=[invsigma;output_invsigma];
accept=[accept;output_acpt];
end;
```

```
% Separate results from two chains into two matrices;
beta_chain1=beta(1:5*N+5,:);beta_chain2=beta(5*N+6:10*N+10,:);
mu_chain1=mu(1:N+1,:);mu_chain2=mu(N+2:2*N+2,:);
invsigma_chain1=invsigma(1:6*N+6,:);invsigma_chain2=invsigma(6*N+7:12*N+12,:);
% Stack  two matrices for two chains into a 3-dim array;
result_beta=cat(3,beta_chain1,beta_chain2);
result_mu=cat(3,mu_chain1,mu_chain2);
result_invsigma=cat(3,invsigma_chain1,invsigma_chain2);

% Create 5 arrays for each of the 5 betas;
result_beta1=result_beta(1:5:end,:,:);result_beta2=result_beta(2:5:end,:,:);
result_beta3=result_beta(3:5:end,:,:);result_beta4=result_beta(4:5:end,:,:);
result_beta5=result_beta(5:5:end,:,:);

% Show the acceptance rate for the two chains for 5 Beta vectors;
accept=accept(2:3,:)

% Calculate convergence diagnostics "mpsrf";
conv_mu=mpsrf(result_mu)
conv_beta1=mpsrf(result_beta1)
conv_beta2=mpsrf(result_beta2)
conv_beta3=mpsrf(result_beta3)
conv_beta4=mpsrf(result_beta4)
conv_beta5=mpsrf(result_beta5)
conv_invsigma=mpsrf(result_invsigma)

% calculate autocorr for beta and mu for each chain
acorr_beta1_chain1=acorr(result_beta1(:,:,1),400);
acorr_beta1_chain2=acorr(result_beta1(:,:,2),400);
…
acorr_beta5_chain1=acorr(result_beta5(:,:,1),400);
acorr_beta5_chain2=acorr(result_beta5(:,:,2),400);
acorr_mu_chain1=acorr(result_mu(:,:,1),400);
 acorr_mu_chain2=acorr(result_mu(:,:,2),400);

% Plot auto-correlation plots;
plot(acorr_beta1_chain1)
figure
plot(acorr_beta1_chain2)
…
plot(acorr_beta5_chain1)
figure
plot(acorr_beta5_chain2)
figure
plot(acorr_mu_chain1)
figure
```

```
plot(acorr_mu_chain2)
figure


% Burn-in and Thinning;
beta1_thin1=thin(result_beta1(:,:,1),5000,50);
beta1_thin2=thin(result_beta1(:,:,2),5000,50);
…
beta5_thin1=thin(result_beta5(:,:,1),5000,50);
beta5_thin2=thin(result_beta5(:,:,2),5000,50);
mu_thin1=thin(result_mu(:,:,1),5000,50);
mu_thin2=thin(result_mu(:,:,2),5000,50);

% Combine all thinned chains into one matrix for each vector;
beta1_allchain=[beta1_thin1;beta1_thin2]; beta2_allchain=[beta2_thin1;beta2_thin2];
beta3_allchain=[beta3_thin1;beta3_thin2]; beta4_allchain=[beta4_thin1;beta4_thin2];
beta5_allchain=[beta5_thin1;beta5_thin2];
mu_allchain=[mu_thin1;mu_thin2];

% Check if auto corr after thinning has reduced to close 0;
acorr_beta1=acorr(beta1_allchain,400);
…
acorr_beta5=acorr(beta5_allchain,400);
acorr_mu=acorr(mu_allchain,400);

plot(acorr_beta1)
figure
…
plot(acorr_beta5)
figure
plot(acorr_mu)

% Calculates summary stats for the params and phat.
% Mean, Std and percentiles for beta and mu vectors;
mean_beta1=mean(beta1_allchain);percentile_beta1=prctile(beta1_allchain,[2.5 5 50 95
97.5]);
std_beta1=std(beta1_allchain); median_beta1=median(beta1_allchain);
…
mean_beta5=mean(beta5_allchain);percentile_beta5=prctile(beta5_allchain,[2.5 5 50 95
97.5]);
std_beta5=std(beta5_allchain); median_beta5=median(beta5_allchain);
mean_mu=mean(mu_allchain);percentile_mu=prctile(mu_allchain,[2.5 5 50 95 97.5]);
std_mu=std(mu_allchain); median_mu=median(mu_allchain);

summary_beta1=[mean_beta1;std_beta1;percentile_beta1]
…
```

```
summary_beta5=[mean_beta5;std_beta5;percentile_beta5]
summary_mu=[mean_mu;std_mu;percentile_mu]

%phat calculation;
obs_start_1=[1];        % position of starting obs for group 1;
obs_end_1=[548];      % position of ending obs for group 1;
…
obs_start_5=[2976];    %location of starting obs for group 5
obs_end_5=[3337];      %location of ending obs for group 5

Data_1=Data(obs_start_1:obs_end_1,:);
…
Data_5=Data(obs_start_5:obs_end_5,:);

% Group 1 estimates;
p1_estimate_1=exp(Data_1(:,1:6)*mean_beta1')./(1+exp(Data_1(:,1:6)*mean_beta1'));
%plug-in-mean-beta approach
p1_estimate_2=mean(exp(Data_1(:,1:6)*beta1_allchain')./(ones(size(Data_1,1),size(beta
1_allchain,1))+exp(Data_1(:,1:6)*beta1_allchain')),2);%calculate phat in each simu, and
use the average;
p1_estimate_3=exp(Data_1(:,1:6)*median_beta1')./(1+exp(Data_1(:,1:6)*median_beta1')
);%plug-in-median-beta approach;
p1_estimate_4=median(exp(Data_1(:,1:6)*beta1_allchain')./(ones(size(Data_1,1),size(bet
a1_allchain,1))+exp(Data_1(:,1:6)*beta1_allchain')),2);%calculate phat in each simu, and
use the median;
…

%Group 5 estimates;
p5_estimate_1=exp(Data_5(:,1:6)*mean_beta5')./(1+exp(Data_5(:,1:6)*mean_beta5'));
p5_estimate_2=mean(exp(Data_5(:,1:6)*beta5_allchain')./(ones(size(Data_5,1),size(beta
5_allchain,1))+exp(Data_5(:,1:6)*beta5_allchain')),2);
p5_estimate_3=exp(Data_5(:,1:6)*median_beta5')./(1+exp(Data_5(:,1:6)*median_beta5')
);
p5_estimate_4=median(exp(Data_5(:,1:6)*beta5_allchain')./(ones(size(Data_5,1),size(bet
a5_allchain,1))+exp(Data_5(:,1:6)*beta5_allchain')),2);

p1=[p1_estimate_1,p1_estimate_2,p1_estimate_3,p1_estimate_4];
p2=[p2_estimate_1,p2_estimate_2,p2_estimate_3,p2_estimate_4];
p3=[p3_estimate_1,p3_estimate_2,p3_estimate_3,p3_estimate_4];
p4=[p4_estimate_1,p4_estimate_2,p4_estimate_3,p4_estimate_4];
p5=[p5_estimate_1,p5_estimate_2,p5_estimate_3,p5_estimate_4];

% output predicted probability to an excel file;
phat=[p1;p2;p3;p4;p5];
xlswrite('E:\Amy\Research\2005 summer\Eric\hudy 2005\phat_matlab.xls', phat)
```

% Function for the Gibbs-Metropolis simulation;

```
function [result_beta,result_mu,result_invsigma,result_acpt]=main_function_estimate(N,
param_beta,param_mu,param_invsigma)

global Data;
% Input data matrix;
Data=[…]

obs_start_1=[1];        % position of starting obs for group 1;
obs_end_1=[548];      % position of ending obs for group 1;
…
obs_start_5=[2976];    %location of starting obs for group 5
obs_end_5=[3337];      %location of ending obs for group 5


sigma_beta1=[…]  %variance matrix for proposal distribution for group 1 (beta1);
…
sigma_beta5=[...]  %variance matrix for proposal distribution for group 5 (beta5);

% Simulate beta11 vector, beta2 vector,... beta5 vector one at a time;
% Simulate mu and sigma from the full conditional MVN and Wishart distributions
directly;

% Initialization;
accept=[0 0 0 0 0]; %initialize
acpt=[0 0 0 0 0];
result_mu=param_mu;
result_beta=param_beta;
result_invsigma=param_invsigma;

for i=1:N
   beta1=mvnrnd(param_beta(1,:),sigma_beta1);
alpha1=full_conditional_beta(obs_start_1,obs_end_1,beta1,param_beta(1,:),param_mu,p
aram_invsigma);
   u1=unifrnd(0,1);
   if u1 <= min(1, alpha1)
      param_beta(1,:)= beta1;
      acpt(1)=1; accept=[accept;acpt];
      acpt=[0 0 0 0 0];
   end;
   %the above code generate beta1 vector and decide if the chain moves to a new vector
or stay.

…..
```

```
  beta5=mvnrnd(param_beta(5,:),sigma_beta5);

  alpha5=full_conditional_beta(obs_start_5,obs_end_5,beta5,param_beta(5,:),param_mu,
param_invsigma)
 u5=unifrnd(0,1);
    if u5 <= min(1, alpha5)
        param_beta(5,:)= beta5;
        acpt(5)=1; accept=[accept;acpt];
        acpt=[0 0 0 0 0];
    end;

param_mu=full_conditional_mu(param_beta, param_invsigma);
param_invsigma=full_conditional_sigma(param_beta,param_mu);
result_beta=[result_beta;param_beta];
result_mu=[result_mu;param_mu];
result_invsigma=[result_invsigma;param_invsigma];
end

result_acpt=sum(accept)/N;
```

% Function of full-conditional of beta;

```
function
[ratio]=full_conditional_beta(obs_start,obs_end,beta_new,beta_old,mu,invsigma);
global Data;

data=Data(obs_start:obs_end,:);
n=size(data,1);
xb_new=data(:,1:6)*beta_new';
tot_yxb_new=sum(data(:,7).*xb_new);
tot_log_new=sum(log(ones(n,1)+exp(xb_new)));
logsum_new=tot_yxb_new-tot_log_new-0.5*(beta_new'-mu')'*invsigma*(beta_new'-
mu');

xb_old=data(:,1:6)*beta_old';
tot_yxb_old=sum(data(:,7).*xb_old);
tot_log_old=sum(log(ones(n,1)+exp(xb_old)));
logsum_old=tot_yxb_old-tot_log_old-0.5*(beta_old'-mu')'*invsigma*(beta_old'-mu');

ratio=exp(logsum_new-logsum_old);
```

% Function of full-conditional of mu;

```
function [draw]=full_conditional_mu(beta,invsigma);

alpha=[…]    %specify hyperprior mean;
omiga=[…]   %specify hyperprior variance;
no_cluster=5;
beta1=beta(1,:); beta2=beta(2,:);beta3=beta(3,:);beta4=beta(4,:);beta5=beta(5,:);

mean_vector=inv(no_cluster*invsigma+inv(omiga))*(beta1*invsigma+beta2*invsigma+
beta3*invsigma+beta4*invsigma+beta5*invsigma+alpha*inv(omiga))';
mean_mu=mean_vector';
cov=inv(no_cluster*invsigma+inv(omiga));
draw=mvnrnd(mean_mu,cov);
```

% Function of full-conditional of sigma;

```
function [draw]=full_conditional_sigma(beta, mu);

R=[…];     %specify scale matrix for sigma (precision) for wishart distribution;
no_cluster=5;
df_prior=7;
beta1=beta(1,:)'; beta2=beta(2,:)';beta3=beta(3,:)';beta4=beta(4,:)';beta5=beta(5,:)';
mu=mu';
beta_mu=(beta1-mu)*(beta1-mu)'+(beta2-mu)*(beta2-mu)'+(beta3-mu)*(beta3-
mu)'+(beta4-mu)*(beta4-mu)'+(beta5-mu)*(beta5-mu)';
scale=inv(inv(R)+beta_mu);
df_post=no_cluster+df_prior;
draw=wishrnd(scale,df_post);
```

Appendix B: Variable selection results given the 6-cluster solution

	Stepwise selection (4 variables)			Stepwise selection (all variables)		
	Parameter	**Estimate**	**s.e**	**Parameter**	**Estimate**	**s.e**
Cluster 1	Intercept	2.61	0.31	Intercept	6.77	0.84
	Elevation	-1.96	0.21	Elevation	-2.53	0.26
	Road density	0.36	0.14	Acid deposition	3.52	0.66
				Mixed forest 2	-0.77	0.19
				Industrial 2	0.45	0.17
				Population density 2	0.30	0.11
Cluster 2	Intercept	-2.54	0.24	Intercept	-2.66	0.29
	Elevation	-1.24	0.29	Elevation	-1.19	0.29
	Road density	0.31	0.15	Evergreen 2	-0.59	0.15
	Total forest	-0.42	0.13	Total forest	-0.49	0.11
	Agriculture	0.51	0.13	Agriculture	0.21	0.10
				Shrubland	-1.28	0.54
Cluster 3	Intercept	-2.39	0.45	Intercept	-4.20	0.61
	Elevation	-3.19	1.00	Elevation	-6.07	1.09
	Road density	0.89	0.41	Evergreen 2	-1.13	0.34
	Agriculture	0.93	0.28	Row_crop 2	0.58	0.26
Clsuter 4	Intercept	0.56	0.36	Intercept	-2.75	0.81
	Elevation	-2.71	0.45	Elevation	-3.39	0.52
	Agriculture	0.75	0.19	Acid deposition	-1.13	0.38
				Mixed forest 2	-0.85	0.32
				Agriculture	0.70	0.21
				Wetlands	-3.78	1.03
Cluster 5	Intercept	-1.81	0.19	Intercept	-1.88	0.21
	Elevation	-2.83	0.31	Elevation	-2.69	0.32
	Road density	0.80	0.19	Road density	0.68	0.31
	Agriculture	0.85	0.16	Evergreen 2	-0.35	0.12
				cs_sHigh_res	-1.82	0.49
				Agriculture	0.67	0.17
				Population density 2	1.52	0.42
Cluster 6	Intercept	-4.98	0.77	Intercept	-5.24	0.85
	Elevation	-1.29	0.61	Elevation	-3.01	0.79
	Road density	0.96	0.22	Acid deposition	0.90	0.34
				Evergreen 2	0.97	0.35
				Population density 2	0.73	0.16

Note: Highly correlated variables are eliminated before variable selection.

Appendix C: AUC measures for the two models after variable selection

	AUC		
	Stepwise (4 variables)	Stepwise (all variable)	4-variable model
Cluster 1	0.82	0.86	0.82
Cluster 2	0.86	0.87	0.86
Cluster 3	0.94	0.95	0.94
Cluster 4	0.88	0.92	0.88
Cluster 5	0.91	0.92	0.91
Cluster 6	0.84	0.86	0.84

Vita

Huizi Zhang was born in Shaoyang City, Hunan Province, China. She graduated from No. 2 High School of Shaoyang City in 1989. She graduated from Beijing University in 1999 with a Bachelor's Degree in Economics.

She began her graduate study at Virginia Polytechnic Institute and State University in the fall of 2001. After receiving a Master of Science degree in Statistics in December, 2002, she continued her study in the Ph.D. program and was awarded the Ph.D. degree in Statistics in August, 2006.

She is a member of the Mu Sigma Rho honorary society in statistics and she served as the treasurer of the society during the 2003 academic year. She is a member of the Virginia Academy of Science and a member of the American Statistical Association.