Clustering Response-Stressor Relationships in Ecological Studies

Feng Gao

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

Doctor of Philosophy
in
Statistics

Dr. Eric P. Smith, Chair
Dr. Samantha C. Bates Prins, Co-Chair
Dr. Dan Spitzner
Dr. George Terrell

June 20, 2007
Blacksburg, Virginia

Keywords: Model based clustering, Voronoi diagrams, BCART, RDA, CCA, model selection

Clustering Response-Stressor Relationships in Ecological Studies

Feng Gao

(ABSTRACT)

This research is motivated by an issue frequently encountered in water quality monitoring and ecological assessment. One concern for researchers and watershed resource managers is how the biological community in a watershed is affected by human activities. The conventional single model approach based on regression and logistic regression usually fails to adequately model the relationship between biological responses and environmental stressors since the study samples are collected over a large spatial region and the response-stressor relationships are usually weak in this situation. In this dissertation, we propose two alternative modeling approaches to partition the whole region of study into disjoint subregions and model the response-stressor relationships within subregions simultaneously. In our examples, these modeling approaches found stronger relationships within subregions and should help the resource managers improve impairment assessment and decision making.

The first approach is an adjusted Bayesian classification and regression tree (ABCART). It is based on the Bayesian classification and regression tree approach (BCART) and is modified to accommodate spatial partitions in ecological studies. The second approach is a Voronoi diagram based partition approach. This approach uses the Voronoi diagram technique to randomly partition the whole region into subregions with predetermined minimum sample size. The optimal partition/cluster is selected by Monte Carlo simulation. We propose several model selection criteria for optimal partitioning and modeling according to the nature of the study and extend it to multivariate analysis to find the underlying structure of response-stressor relationships. We also propose a multivariate hotspot detection approach (MHDM) to find the region where the response-stressor relationship is the strongest according to an R-square-like criterion. Several sets

of ecological data are studied in this dissertation to illustrate the implementation of the above partition modeling approaches. The findings from these studies are consistent with other studies.

Dedication

To my parents and my lovely daughter, Cindy.

Acknowledgements

I would like to express my deepest appreciation to my advisor Dr. Eric Smith for his guidance and support. Throughout my doctoral work he encouraged me to develop independent thinking and research skills. He continually stimulated creative thinking and greatly assisted me with scientific writing. His academic insight and personal charm have been invaluable to my career development. I am also very grateful for my co-advisor Dr. Samantha C. Bates Prins. I would like to thank her for all the technical help, generous time she gave me, and patience for my research. It is not possible to build the big framework of the Monte Carlo simulation without her support. I am extremely thankful for having an exceptional doctoral committee and wish to thank Dr. Dan Spitzner and Dr. George Terrell for all the thoughtful ideas, valuable comments and suggestions they shared with me. Thanks also go to Dr. Keying Ye for his encouragement and inspiration inside and outside the classroom which kept me moving forward and going though ups and downs. I would also like to thank my internship mentor from CapitalOne, Dr. George Paslaski for his guidance and insightful ideas which lead me to the exploration of Bayesian regression and classification trees.

I also want to extend my thanks to my old friends here in Virginia Tech, especially Younan, Huizi, Raina, Kim, Brooke, Zhengrong, David, Mihyun, Ying, Bing, Li, Wen and Xiaowei etc. and my dancing group friends Joy, Chunhong, Xuejuan, Sally and Ling. Thanks also go to Peter, Guangzhen and Patrick for the spiritual inspiration. Thanks also to the many unmentioned for so many reasons that are too long to name. They fulfilled my life as a Ph.D student at Virginia Tech.

This research was partially funded by EPA Grant RD-83136810-0: Model-based clustering for classifications of aquatic systems & diagnosis of stress, which I would also like to show my appreciation to.

Special thanks go to my husband Shiming and daughter Cindy for their support over all these years. I could not imagine that I can go this far without them.

Contents

List of Figures

List of Tables

1 Introduction

In ecological studies, it is important to find the statistical relationship between biological responses and external (environmental) stressors (Yuan & Norton, 2004). Biological responses refer to the variable(s) used to measure biological status, abundance or functions of abundance in a watershed of interest. The responses could be metrics/indexes of species, species counts, or the status of species (e.g. present or absent). They are chosen to be the indicators of the health or quality of the watershed. Stressors are variables used to measure the environmental variation or human perturbations in a watershed of interest and they are the possible causes of watershed impairment. The response-stressor relationship could help resource managers/researchers to find the possible causes of watershed degradation and take corresponding actions to prevent them. Most often, for a large spatially balanced dataset, the pattern of these relationships is only locally significant. Thus, these relationships can not be detected using conventional statistical analyses such as regression analyses and multivariate methods directly. As a result, clustering based on response-stressor relationships is proposed as a solution and will be referred to as model-based partition/clustering.

Many model-based clustering methods have been investigated by researchers in the medical, aerospace and marketing areas. In general, these methods can be divided into two categories: One category involves probability modeling, such as the finite mixture model method or the latent variable regression approach. This modeling approach assigns each observation into a group based on the posterior probability of a latent group variable and the partition is spatially overlapping. The other category involves non-parametric modeling using as tree-based approaches. Two methods are Classification and Regression Trees -CART (Breiman *et al.,*1984) and Bayesian Classification and Regression Trees -BCART (Chipman, George and McCulloch, 1998). These approaches generally partition the observations over the predictor space of stressor variables.

In this study, we are especially interested in partitioning a geographical/spatial region into clusters with a fitted statistical model in each cluster. The BCART method can be used to achieve this goal by separating the splitting and modeling variables of the model. However, a geographical/spatial region usually has an irregular shape while the BCART method has been only applicable for a domain with a rectangular shape. In this dissertation, we propose an enhancement of the BCART method that uses coordinate transformations of spatial splitting variables from an irregular domain to a rectangular domain. This method is called the adjusted BCART approach. We then applied this approach to brook trout data.

The Voronoi diagram-based partition method can also be used to accommodate the needs of spatial partitioning. Holmes *et al.* (2001) proposed a Bayesian partition model using the Voronoi diagram. Because of the nature of the Bayesian approach, it is hard to deal with high dimensional multivariate analysis using this method. In this dissertation, we propose a non-Bayesian spatial partition modeling approach based on a Voronoi diagram technique as the alternative. This approach has some advantages over the Bayesian partition model and is flexible enough for different statistical analysis including multivariate situations. We call this approach as the spatial partition model for short when we partition on latitude and longitude. "Model" in the case of partition modeling has two components: the partition/clustering model and the statistical model within each cluster. The goal of model selection in spatial partition modeling is to find the correct cluster(s) within the geographical region of interest and the correct parametric/nonparametric model within the cluster(s). Our research focuses on finding the underlying structure of the region or detecting the specific region with a response-stressor relationship of interest. In general, partition model performance should be measured by aspects such as the correctness of number of clusters, the misclassification rate given the right number of clusters, the accuracy of prediction or adjusted R-square for the regression model, etc. The model selection criterion used will also affect the final model performance.

This dissertation is organized seven chapters. In Chapter two, we give a review of the leading literature on the various partition methods mentioned above with focus on Bayesian CART, spatial partition modeling approaches and model selection approaches. In Chapter 3, we extend and enhance the BCART method by using a coordinate transformation for spatial variables and compare its application on a brook trout dataset along the east coast with the result of the spatial partition modeling using the Voronoi diagram-based method. The data was analyzed by Zhang *et al.* (2008). In Chapter 4, we extend the spatial partition modeling approach with a logistic model to one with a multicategory logit model. In Chapter 5, the spatial partition modeling is extended to multivariate analysis methods such as canonical correspondence analysis (CCA) and redundancy analysis (RDA). We propose a Refined Multivariate Partition Modeling (RMPM) approach using a BIC-like criterion to find the underlying multivariate structure according to the multivariate response-stressor relationship. In Chapter 6, we focus on detecting the hotspot of the strongest multivariate response-stressor relationship. A Multivariate Hotspot Detection Modeling (MHDM) approach is presented to find the region with strongest response-stressor relationship in terms of R-square-like values. A stressor selection algorithm is incorporated into the partition process. We applied RMPM and MHDM approaches to West Virginia data. Finally, in Chapter 7, we summarize the properties of the non-Bayesian partition modeling, draw conclusions, and propose our thoughts on future research.

2 Partition Modeling Approaches and Model Selection Criteria

In practice, it is very important for researchers and managers to have a better understanding of the watershed ecosystem they are monitoring. The watershed ecosystem consists of the biological community and environmental characteristics such as chemistry, landscape, hydrological and other human perturbation factors which we call stressors. Usually, the relationship between the biological response(s) and environmental stressors changes when measured over a range of spatial scale and typically the strength of the relationship is often weak over a large region. The clustering or partition modeling approaches are structured around changes in the biological community as a function of changes in stressors and aim to find strong response-stressors relationship by finding strong local patterns. By combining these patterns, resource managers should be able to better diagnose ecological stress and impairment. It can also help them with the design of efficient monitoring strategies for stress evaluation as required by the Clean Water Act (Smith, 2003). According to the Clean Water Act, agencies need to assess their surface water and report the condition of those waters periodically.

There are two types of clustering schemes in the literature. One is variable-based clustering and the other is model-based clustering. The former is usually called cluster analysis. The goal of variable-based clustering is to find an optimal grouping for which the observations within each cluster are similar but are different between groups. Clusters are formed based on a measure of distance or similarity/dissimilarity between groups. K-means clustering, hierarchical agglomerative methods and regression trees are based on this mechanism.

The model-based clustering method will choose clusters such that the model structure under the class assignment is similar. The model within each cluster can be any parametric or non-parametric statistical model, eg. a multiple regression model, logistic regression model or multivariate model etc. Since we are interested in finding the

relationship between a response(s) and stressors over a large spatial region, the model-based clustering method is our research focus. In this chapter, three model-based clustering methods will be briefly presented and compared and then the methods for model selection and evaluation will be introduced.

Over the past 20 years, several model-based methods have been proposed and developed. The most often mentioned approach in the literature is the finite mixture modeling approach. Actually, it has been shown (Banfield & Raftery, 1993) that various heuristic methods of cluster analysis such as K-means clustering and hierarchical agglomerative are the special case of this mixture approach when there are no stressors. This method will be briefly introduced in Section 2.2. Bayesian Classification and Regression Tree (BCART) is a relatively new classification modeling approach proposed by Chipman (Chipman *et al.*, 1998). It has been used in marketing studies to predict credit card usage effectively (Gao, 2005). In this chapter, it will be introduced after the Classification and Regression Tree (CART) (Breiman *et al.*, 1984) model since it is an extension of the CART model. BCART modeling inherits some properties from the CART approach. One important property is that the BCART analysis is invariant to the monotonic transformation of splitting (explanatory) variables (De'ath & Fabricius, 2000). We call CART and BCART tree-based approaches. The Voronoi-based partition modeling approach is another model-based clustering approach proposed recently (Bates Prins *et al*., 2006). It is very useful for clustering and modeling over a large spatial region and the method fits our interest very well. The details of this method will be introduced in Section 2.3. Model selection and evaluation is an important component in the development of partition modeling approaches, we will discuss some methodology in the last section of this chapter.

2.1 Tree-based partition approaches

The tree-based methods use a binary tree structure to classify the observations into disjoint groups. The idea dates back to the recursive partitioning ideas of Morgan and Sonquist (1963). Breiman *et al.* (1984) set up the foundation for current approaches

to tree-based modeling. Chipman *et al.* (1998) and Denison *et al.* (1998 & 2001) extended this idea to Bayesian tree modeling.

2.1.1 Classification and regression trees (CART)

Classification and Regression Trees (CART), introduced by Breiman *et al.* (1984) is a method to model a response variable (y) which is either categorical (classification trees) or numeric (regression trees) as a function of explanatory variables, $\mathbf{x}$. In our application, y represents the status of stream sites (extirpated or present) and $\mathbf{x}$ is a collection of landscape and other variables. The relationship between y and $\mathbf{x}$ is constructed by repeatedly splitting the explanatory variable(s) $\mathbf{x}$ according to rules to form a binary tree **T** and evaluating the model for the subset of data defined by each split. At each split the data is partitioned into two distinct, non-overlapping groups, each of which is made as homogeneous (in the sense of the model) as possible. The splitting could continue until each group size was small with an additional desire to keep the tree reasonably small. The quality of the tree is determined by a numerical criterion which has a penalty to keep the tree from being too large. Splitting is usually continued until an overly large tree is grown. This generates a sequence of trees, each of which is an extension of the previous tree. A 'best' tree is selected by pruning the largest tree according to a model selection criterion such as cost-complexity or crossvalidation. When the splitting and pruning stops, tree **T** has b terminal nodes with each terminal node corresponding to a region in the data space. The b regions corresponding to the b terminal nodes are disjoint. The key problems in this process are: (1) How to split; and (2) How to find the best tree.

According to De'ath, and Fabricius (2000) each group in the terminal nodes of the selected tree is typically characterized by either the distribution (categorical response) or mean value (numeric response) of the response variable, group size, and the values of the explanatory variables. The homogeneity of nodes is measured by impurity, which takes the value zero for completely homogeneous nodes, and increases when the homogeneity decreases. Therefore, maximizing the homogeneity of the groups has the same meaning

as minimizing their impurity (Breiman, 1984). Several impurity criteria were suggested by Breiman. In fact, these impurity minimizing criterions can be viewed as a misclassification rate for the classification tree and the mean square error for the regression tree. Accordingly, an overlarge tree can be formed through minimizing impurity.

CART analysis has many advantages: First it makes no assumptions regarding the underlying distribution of values of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal. Second, since only the rank of a numeric explanatory variable determines a split, trees are invariant to monotonic transformations of numeric explanatory variables. Third, CART can also deal with missing values even when important predictor variables are not known for all sites. Fourth, CART requires relatively little input from the analyst as it is a relatively automatic "machine learning" method. Finally, CART trees are relatively simple for non-statisticians to interpret because of its simple logic.

Despite its many advantages, there are many disadvantages of CART. First, because each split depends on the previous split, the CART solution is a local optimal solution. Second, the distributions in the terminal nodes are all degenerate. All points in a single terminal node are classified as either the most abundant category or as the mean values of the data points associated with the node. Third, CART uses a 'greedy algorithm' to produce a sequence of trees, all of which are refinements of the previous tree in the sequence. Thus, the whole tree space is not explored.

2.1.2 Bayesian tree models

The Bayesian tree model overcomes some uncertainties and limitations of the CART model. Essentially, the conventional CART model was proposed as a data mining tool rather than a modeling tool. The Bayesian approach improves the conventional CART model by allowing for more sophisticated models (such as linear regression) in a conventional tree to replace the terminal node means or class probabilities. In addition, the Bayesian approach induces a posterior distribution to guide a stochastic search

towards more 'promising' tree models by fully exploring the model space, while the conventional CART model uses impurity reduction for tree construction resulting in limited set of trees. Because the posterior is based on tree structure and terminal node models, the form of terminal node models become an integral part of the stochastic search. It is noted that there are two key elements in Bayesian tree modeling. One is the introduction of prior distributions for all unknown parameters associated with the tree structures and terminal node models. The other is the stochastic search for interesting models by the Markov chain Monte Carlo method.

One important thing that makes the application of this Bayesian CART approach outstanding in our ecological study is that we can use one subset of explanatory variables **x** to define the binary tree while using another subset to fit the terminal models (e.g. regression models). In our application, the x set of variables consists of landscape variables, other factors and site locations. The site locations (latitude and longitude) may be used as the splitting variables and the other variables may be used as the explanatory variables for the regression models. Those two subsets can be disjoint. This due to the fact that the tree priors and the terminal model priors can be set up separately. The recent implementation of BCART to regional-scale eutrophication models (Lamon & Stow, 2004) used latitude and longitude as splitting variables and others as modeling variables. Since the splitting on these two variables is binary and only one at a time, the shapes of the partitioned spatial regions are all rectangles. (The Voronoi diagram-based partition approach gives much more flexibility with regard to the shape of regions partitioned and thus has advantages in finding cluster with better fits.). Similarly, BCART gives more flexibility in modeling within terminal nodes and the tree. This method has been extended to generalized linear models by Chipman *et al.* (2003) which, for example, allows us to use logistic regression modeling at terminal nodes.

2.2 Finite mixture model approaches

Finite mixture models represent an alternative approach for grouping observations into unobserved clusters/segments. The models have been widely used in applications from biology and medicine to physics, economics, and marketing. This method can be

applied to data that are obtained from various groups, of which the affiliations are not known. It is currently a hot topic in the field of model-based clustering method because the distribution of observations can be modeled as multi-modal distributions. The most often explored finite mixture model in the literature of classification and modeling is the mixture of linear models which was named as clusterwise regression (DeSarbo & Cron, 1988) or latent class regression (Magidson *et al*., 2003). The model has been extended to the clusterwise general linear model (Wedel & DeSarbo, 1995). Fraley and Raftery (2002) provided a very good review of the general methodology. McLachlan and Peel (2000) provided comprehensive details of the model in the book 'Finite Mixture Models'. Finite mixture models with a fixed number of components are usually estimated with the EM algorithm within a maximum likelihood framework and with MCMC sampling within a Bayesian framework (Leisch, 2004). Since we are focusing on linear regression modeling within each cluster, we will only look at mixtures of linear models. This modeling approach will simultaneously find clusters and build linear regression models within the clusters.

Suppose $\mathbf{y}$ is a (possibly multivariate) dependent variable with conditional density, *f*, $\mathbf{x}$ is a vector of independent variables, π_k is the prior probability of cluster/component *k*, $\boldsymbol{\theta}_\mathbf{k}$ is the component specific parameter vector for the component specific density function f_k, *K* is the number of cluster/components, and $\boldsymbol{\psi} = (\pi_1, ..., \pi_K, \boldsymbol{\theta}_\mathbf{1}', ..., \boldsymbol{\theta}_\mathbf{K}')'$ is the vector of all parameters. The conditional mixture density function for the i^{th} observation y_i of $\mathbf{y}$ is $h(y_i \mid x_i, \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(y_i \mid x_i, \boldsymbol{\theta}_\mathbf{k})$ with $\sum_{k=1}^{K} \pi_k = 1$ and $0 \le \pi_k \le 1$.

The posterior probabilities can be used to classify data by assigning each observation to the most likely class. Thus, the cluster the observation is assigned to is given by the cluster with the maximum posterior probability. That is, we find j that maximizes

$$P(j \mid x_i, y_i, \boldsymbol{\psi}) = \frac{\pi_j f(y_i \mid x_i, \boldsymbol{\theta}_j)}{\sum_k \pi_k f(y_i \mid x_i, \boldsymbol{\theta}_k)}, \; j = 1, 2, \ldots K$$

The EM algorithm iterates between estimating posterior probabilities and maximizing the likelihood according to those probabilities. Estimates of the posterior probabilities for each observation are given by

$$\hat{p}_{ij} = P(j \mid x_i, y_i, \hat{\boldsymbol{\psi}}) = \frac{\hat{\pi}_j f(y_i \mid x_i, \hat{\boldsymbol{\theta}}_j)}{\sum_k \hat{\pi}_k f(y_i \mid x_i, \hat{\boldsymbol{\theta}}_{\mathbf{k}})} \qquad \hat{\pi}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_{ij}$$

where N is the total number of observations. Given the $\hat{\pi}_k$, we now maximize the log-likelihood for each component separately using the posterior probabilities as weight, i.e. $\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{N} \hat{p}_{ij} \log f(y_i \mid x_i, \boldsymbol{\theta}_j)$. The E- and M- steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached**.**

Usually in applications, the number of clusters is unknown and has to be inferred from the data, along with the parameters in the cluster densities. The most frequently used method to determine the number of clusters is an information-based criterion (such as BIC) which will be discussed later in model selection. Fraley and Raftery (2002) has provided details on how BIC works in choosing the right cluster. The finite mixture modeling approach usually gives us spatially overlapping clusters. It is an alternative tool to analyze ecological data.

2.3 Voronoi-diagram based partition model approaches

The analysis of spatial data is of profound importance in our research in environmental ecology. In general, we are interested in investigating the presence of clusters with different response-stressor relationships within different disjoint regions geographically. The motivation behind the Voronoi-diagram based partition model is that points (sites) nearby in some spatial measurement (such as longitude and latitude) have a similar response-stressor relationship, or have the same distribution in the response space.

By relationship, we mean anything that relates responses and stressors such as the correlation, regression etc. The BCART approaches can be used to do this analysis by using spatial variables as splitting variables and other variables as modeling variables within each terminal node, just as Lamon and Stow (2004) did. Although the BCART approach will provide a solution by separating the splitting variables and modeling variables within each terminal node, the hierarchical tree structure embedded inside the approach tends to induce strong correlations in the parameters of the model as the variables are implicitly related to one another through the tree structure. Furthermore, the shape of the spatial region partitioned by terminal nodes is rectangular. To avoid these problems, a Voronoi diagram-based partition approach is used as it has more flexibility. Although the method has similarities with tree structured models in that the data space is partitioned into disjoint regions, the spatial partition model does not embed a hierarchical tree structure, instead, it uses a Voronoi-diagram technique to randomly split the space and finds the 'optimal' groups/clusters through some criteria such as BIC. The following section introduces spatial partition models after a brief review of the Voronoi diagram.

2.3.1 Voronoi diagrams and spatial partitioning

Voronoi diagrams were first discussed by Gustav Peter Lejeune-Dirichlet in 1850 (Lejeune-Dirichlet, 1850). Following a hiatus of more than a half of a century, Voronoi wrote a paper (Voronoi, 1907) about these diagrams and hence the name "Voronoi diagram".

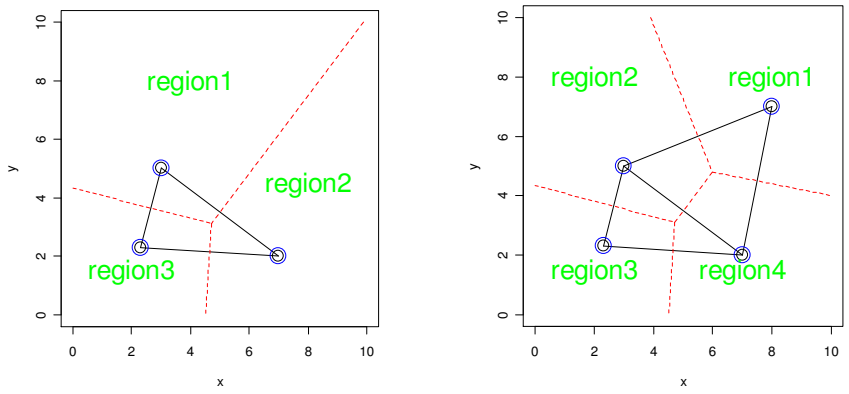
A Voronoi diagram, sometimes called a Dirichlet tessellation, is a method to formulate the disjoint spatial region. It is a geometric structure that represents proximity information about a set of sites or objects. This technique has many applications in areas ranging from archaeology to zoology.

There are a variety of algorithms available to construct Voronoi diagrams. The basic idea is to start from its dual, the Delauney triangulation. Suppose we have a set of nodes (or locations) $o_j = (x_j, y_j), j = 1,...,k$ which are arbitrarily distributed on a two-dimensional region/plane. A triangulation is obtained by connecting those nodes to a set of triangles such that triangle vertices are nodes and no triangle contains nodes other than

its vertices. If there is no node contained in the circumcircle of each triangle, then this triangulation is called the Delauney triangulation (Renka, 1996). After the Delauney triangulation is done, the circumcircle centers are determined. Voronoi cells are convex polygons formed by connecting the circumcircle center points according to the neighborhood relations between the triangles. In other words, the polygon vertices are the circumcenters of those triangles in Delauney triangulation and the polygon sides lie on perpendicular bisectors of the triangulation edges. Thus, the Voronoi diagram creates K spatially disjoint polygons with k nodes such that points in the region/plane closer to one node will be assigned to that polygon containing the node.

Figure 2.1 illustrates this process for 3 and 4 nodes. The solid lines are sides of Delauney triangles and the dotted lines are boundaries of Voronoi diagram. The circles or points are randomly generated group centers for one tessellation. The solid lines form the Delauney triangles and the dashed red lines form the boundary of partitions. The packages 'deldir' and 'tripack' in R can create constrained two dimensional Voronoi tessellations.

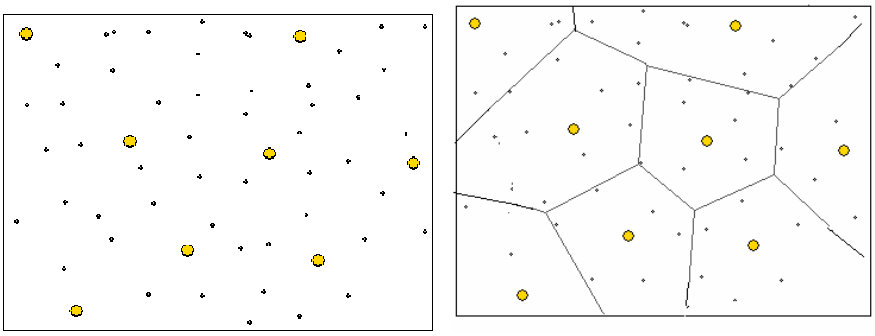
Figure 2.1 Formation of Voronoi diagram, for 3 and 4 cluster partition.



The following is an example on how the Voronoi diagram works for a fixed set of centers (Muhamma,http://www.personal.kent.edu/~rmuhamma/Compgeometry/MyCG/CG/VoroDiagram/vorocli.htm). Suppose we have a few hundred environmental sampling stations

scattered throughout a region and we have eight data collection centers (yellow points in Figure 2.2a). We want to assign each sampling station to the nearest collection center. Using the Voronoi diagram technique, the clustered region is created as shown in Figure 2.2b.

Figure 2.2 An example of Voronoi diagram.



(a) (b)

Since the motivation behind the partition model is that points (sites) nearby in some spatial measurement(s) have a similar relationship between response-stressor, Voronoi diagrams can be used to obtain a solution. In the Voronoi diagrams, one random set of centers can give us one partition, which we call it one tessellation. For a given number of partitions/number of clusters of a partition, we can generate hundreds of thousands of different partitions by using a selecting random set of centers each time. The question is: how do we know which partition is the right partition for the fixed number of partitions? In this research, we propose several model selection criteria for different modeling situations in the later chapters.

2.3.2 Bayesian partition models

The Bayesian partition method was proposed by Holmes *et al.* (2001). It has similarities with the Bayesian tree structured models in that (1) the data space is partitioned into disjoint regions, (2) for any proposed partitioning of the space, the marginal likelihoods of the models can be obtained by using conjugate priors and (3)

Markov chain Monte Carlo simulation techniques are used to find the best partition models. But the Bayesian partition approach uses the Voronoi tessellation technique to randomly split the predictor spaces into the disjoint regions instead of using a hierarchical tree structure to split the predictor spaces into disjoint terminal nodes. The procedure of finding the "best" partition model is very similar to that of the Bayesian tree model approach.

This method has been applied to the analysis of spatial count data in estimation of disease risk surface (Denision & Holmes, 2001). The assumption of the Bayesian partition model approach is that observations from the same cluster have the same distribution, i.e. all clusters are formed from the same distribution family with different parameters and clusters are independent of each other. With a similar assumption, the non-Bayesian spatial partition model can give us better or equally good results and can easily be extended to multivariate data. With reduced assumptions, Voronoi diagram-based nonparametric modeling will work well in some situations where parametric modeling does not work. Besides, the spatial partition method is simpler to understand and easier to interpret.

2.3.3 Voronoi diagram-based partition model

Bates Prins *et al.* (2006) proposed the method of clustering sites to maximize the strength of cluster-wise stressor-response relationships within clusters. This method uses Voronoi diagrams to randomly assign sites to one of the clusters based on some subset of explanatory variables (e.g. latitude and longitude or width and elevation). The method selects the single stressor-response relationship that is optimal and fits the chosen model to all sites in that cluster. An application of the method to data from the Mid-Atlantic Highlands was presented in their paper. In the paper, attention is restricted to the relationship between a single response variable and a single explanatory variable. The method will be extended to multivariate analysis such as CCA, RDA later in Chapter 5. Here only the basic idea of this method is introduced.

Let y be the response variable and $\mathbf{x}$ be the vector of p stressors. We have n observations $(y_i, x_{1i}, ..., x_{pi})$, $i = 1, 2, ...n$ over a large region defined by spatial variables such as latitude and longitude. Each observation was taken from a particular site with the two dimensional spatial measurement variables associated with each observation. The subregions will be formed such that within a subregion/cluster, all observations have a similar dominant predictor-response relationship and between clusters, relationships are possibly different. A Voronoi diagram is used to partition the region defined by the two spatial measurement variables into K non-overlapping subregions/cell. Let n_k be the number of observations placed in the k^{th} subregion with $n_1 + n_2 + ... + n_K = n$. The K cell centers are generated randomly over the entire space for one tessellation.

The random assignment of observations to cells is repeated a large enough number of times, resulting in many possible groupings of the sites. The optimal partition of the region should be found if the number of tessellations is large enough relative to the total number of observations. Optimality is application specific. In general, we want the optimality criterion to be able to find the correct or approximately correct number of partitions, membership of objects within these clusters and models within each cluster. The process to find the optimal partition/tessellation will be affected by many factors such as the interest of the analysis, models within clusters, minimum number of observations in each cluster etc.

2.4 Model selection and assessment

The purpose of model selection is to find the optimal model or a set of models which fit the data best (Lipkovich, 2003). Besides common uses such as the choice of the degree of a polynomial regression and choice among regressions containing different numbers of explanatory variables, model selection is used in the choice of the number of factors in factor analysis and choice of the number of clusters in cluster analysis (Sclove, 1993).

The most important component in the model selection process is the model selection criterion. Optimal values of the criterion will help to determine the "best" model/cluster. Sclove (1993), Hastie, Tibshirani and Friedman (2001) and others have discussed the various aspects of model selection criteria. Penalized log-likelihood is the basis for many important model selection criteria. In the context of clustering, Hawkins *et al.* (2001) used a simulation study to evaluate the influence of some factors such as cluster separation and mixing proportions on the performance of 22 criteria for choice of the number of clusters in mixtures of linear regression models.

2.4.1 Model selection and criteria for number of cluster

The model selection process can be viewed as a learning process. Hastie *et al*., (2001) pointed out that the generalized performance of a learning method relates to its prediction capability on independent test data. Thus, having chosen a final model, the estimation of the prediction error on new data defines the criterion for model assessment. Several methods can be used to select and assess a model simultaneously. When we have enough data, a simple and efficient way is to randomly divide data into two groups, training and test. Then we can use the training set to fit the models and use the test set to estimate the prediction errors for these models. The preferred model will have small prediction error. But in ecological studies, the datasets usually are not large enough. Two alternative methods to do model assessment and selection in this case are (1) compensation of training error when using all data as training data and (2) using crossvalidation and/or bootstrap methods when the training datasets are not large enough.

Let y be the response variable, $\mathbf{x}$ be the explanatory vector, N be the total number of observations in the training data, and $\hat{f}(\mathbf{x})$ be the predicted value from a training data set. A squared loss function for the prediction measurement is $L(y, \hat{f}(\mathbf{x})) = (y - \hat{f}(\mathbf{x}))^2$. The training error is the average loss over the training sample, i.e. $\hat{err}_{train} = 1/N(\sum_{i=1}^{N} L(y_i, \hat{f}(\mathbf{x}_i))^2$. Since the same data is being used to fit the model and assess its error, the training error is less than the true error, the expected prediction error.

Hence it is an overly optimistic estimator of prediction error. The model performance measurement is usually the test error which is the expected prediction error over an independent test sample $err_{test} = E[L(y, \hat{f}(\mathbf{x}))]$. In general, training error constantly decreases with model complexity (increased number of clusters and increased complexity of model within each cluster). The downward bias of training error can be estimated by the following quantity

$$bias = \frac{2}{N}\sum_{i=1}^{N}\text{cov}(f(\mathbf{x}_i), y_i)$$

The true prediction error can be approximated by the important relationship $Err_{in} = E[\hat{err}_{train}] + \frac{2}{N}\sum_{i=1}^{N}\text{cov}(f(\mathbf{x}_i), y_i)$. Here Err_{in} is in-sample error, which is the error in the new response values at each of training point $\mathbf{x_i}$. The bias can be expressed as $bias = \frac{2}{N}\sum_{i=1}^{N}\text{cov}(f(\mathbf{x}_i), y_i) = p\sigma_\varepsilon^{\,2}$, where p is the number of parameters involved in the modeling and $\sigma_\varepsilon^{\,2}$ is the variance of the model's random error. The expression $Err_{in} = E[\hat{err}_{train}] + \frac{2p}{N}\sigma_\varepsilon^2$ is analogous to the C_p statistic in linear regression modeling. The Akaike Information Criterion (Akaike, 1973) $AIC = -2*\ln L + 2p/n$ is a similar estimate of in-sample testing error, where $\ln L$ is the log likelihood of the training data.

In summary, one way to estimate the prediction error is to estimate the bias and then add it to the training error. The approach using AIC or BIC is based on this framework. Another way to estimate the prediction error is to estimate test error directly using crossvalidation and bootstrap. This will be introduced at the end of this chapter. As we recall, CART model selection uses a crossvalidation approach to select its final tree.

2.4.2 BIC-like optimization criteria

The Bayesian information criterion is $BIC = -2*\ln L + p\ln(N)$, where L is the likelihood for data of size N and p is the number of parameters involved in the modeling. When the model is a Gaussian model and assuming $\sigma_\varepsilon^{\,2}$ is known, $-2*\ln L$

is equal to $\sum_i (y_i - \hat{f}(\mathbf{x_i})^2 / \sigma_\varepsilon^2$ and $BIC = \frac{N}{\sigma_\varepsilon^2}(Er\hat{r}_{train} + n_p \ln(N)\frac{\sigma_\varepsilon^2}{N})$, which is proportional to the AIC and tends to penalize more complex models heavily. Thus the BIC usually chooses models that are simpler.

Choosing the model with the minimum BIC is equivalent to choosing the model with largest posterior probability. Raftery (1995) showed that when the baseline model is the null model with no independent variable, the BIC for the current model is

$$BIC = -\chi_m^2 + p\ln(N)$$

where χ_m^2 is the likelihood ratio test statistic for testing the current model m versus the null model.

For Gaussian linear models, the first part of the BIC is related to the residual sum of squares and measures the lack of fit of the current model. The second part is the penalty associated with the number of parameters in the model. The per parameter penalty is $\ln(N)$ for this criterion. The formula indicates that an additional parameter will be included in the model if improvement in the lack of fit is greater than $\ln(N)$.

For one tessellation in the spatial partition model, the BIC-like criterion for the k^{th} cluster/cell within the tessellation can be written as: $BIC_k = n_k \ln(1 - r^2{}_k) + p_k \ln(n_k)$, where the model in cell *k* has r-square given by r_k^2, and the sample size is n_k. K is the number of clusters for this tessellation and $K \geq 2$. One optimal criterion for the tessellation is $BIC = \sum_{k=1}^{K} BIC_k = \sum_{k=1}^{K} n_k \ln(1 - r_k^2) + p_k \sum_{k=1}^{K} \ln(n_k)$. The number of clusters k can be chosen by finding the value of $k = 2,...K$ that optimizes BIC.

2.4.3 V-fold crossvalidation

In the CART model, Breiman *et al.* (1984) used *v-fold* crossvalidation to choose the tree size. Using the same approach, we can also determine the number of clusters in

Voronoi-based modeling. In this process, we divide the data into v mutually exclusive subsets of equal size first. Usually 5 or 10 fold crossvalidation is recommended. Then we drop out each subset in turn, building a tree or Voronoi diagram using the remaining data from which we predict the response for the omitted subset. Then we calculate the estimated prediction error for each subset and sum over all subsets. The above steps are repeated for each possible size of the tree or number of clusters. We will select the tree or tessellation with smallest estimated error rate. The crossvalidation estimates of error often drop rapidly to a minimum then slowly increase. The curve of crossvalidation estimates of error versus tree size or the number of cluster varies for different crossvalidation runs, so a one standard error rule (which is equivalent to Tibshirani's Gap (Tibshirani *et al*., 2001) statistic) was used by Brieman as a tool to choose the best tree i.e. the smallest tree such that its estimated error rate is within one standard error of the minimum.

3 Adjusted BCART, Its Application and Comparison with Voronoi Diagram Based Method

As noted in Chapter 2, BCART is a very useful method for partitioning and modeling. In ecological assessment studies, we want to classify the response-stressor relationships over a large spatial region. BCART can help us achieve this goal by setting up spatial measurement variables as splitting variables to define a tree and using the rest of the explanatory variables as the model fit predictors for terminal node models of the defined tree. However, since the splits on spatial variables are of binary nature, the clustered regions are usually in the shape of a rectangle. This problem hinders its practical use as a partition modeling approach to solve our problem. To deal with this problem, we propose a coordinate transformation method on spatial variables which monotonically transforms the spatial variable from its original domain to a new computational rectangular domain. The CART approach has the property of invariance to the monotonic transformation of splitting (explanatory) variables (De'ath & Fabricius, 2000). The BCART method inherits this property. A non-regular shape from a map can be transformed to a rectangle or square by a coordinate transformation. After incorporating the coordinate transformation, the BCART method is more effective in solving the problem in our situation. In this chapter, we will introduce this method first, apply this method to brook trout data within the eastern United States and then compare the result with the one using the Voronoi-based method by Zhang *et al.* (2008).

3.1 Coordinate transformation for spatial variables

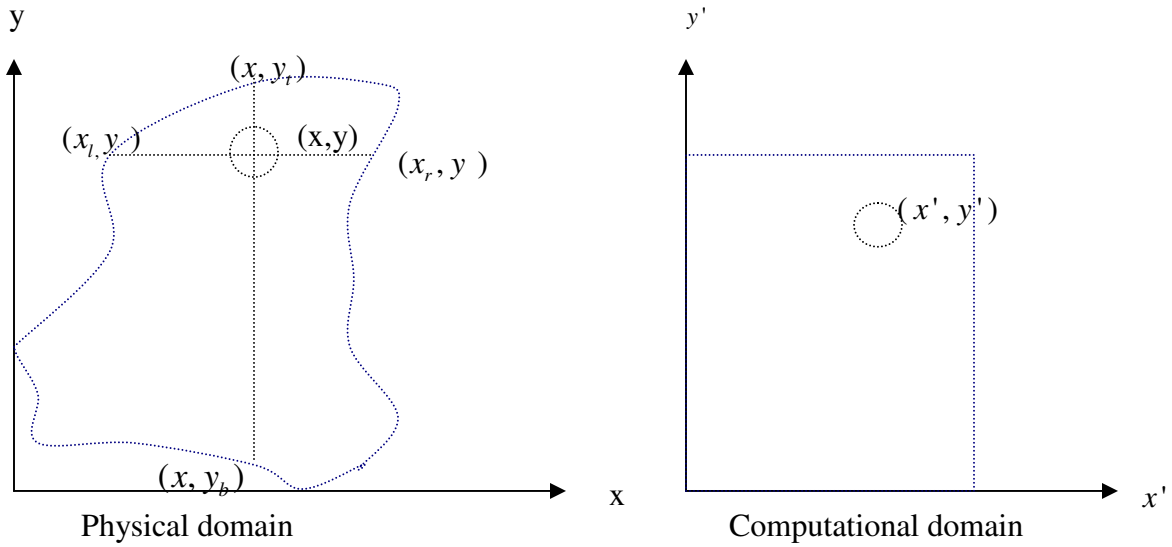
3.1.1 Basic principle

For a two-dimensional non-regular shape, we propose a coordinate transformation method to transform it to a rectangle or square shape. Suppose we have two types of coordinate spaces, one represents the physical space and the other is the desired rectangular domain or the computational space. We can use an algebraic equation to relate the points in the physical domain to those of the computational domain. The mapping can be one-to-one with an appropriate coordinate transformation equation.

The numerical process for this mapping is shown in Figure 3.1. Suppose we know the values of all boundary coordinates. Let (x_l, y) and (x_r, y) be the left and right boundary coordinates corresponding to any point (x, y) within the physical domain and (x, y_t) and (x, y_b) be the top and the bottom boundary coordinates. Left, right, top and bottom boundaries are defined according to the known boundary curve. For any convex boundary shape, we propose the following coordinate transformation relations to transform any point (x, y) of physical domain to corresponding computational domain as

$$x' = \frac{x - x_l}{x_r - x_l}, \; y' = \frac{y - y_b}{y_t - y_b}$$

Figure 3.1 Coordinate transformation from physical domain to computational domain.



In real life, we do not have the specific analytical function for the shape of a physical domain and thus we do not know all the values on the boundary. Many complicated numerical algorithms, known as grid generation techniques (Hoffmann and Chiang, 1993) give a numerical solution by solving a second-order, partial differential equation. The numerical solution of the partial differential equation defines and controls the spatial mappings between the computational domain and the physical domain when only part of the boundary is known. However, the numerical solution of the partial

differential equation usually involves complicated algorithms and cannot guarantee the monotonic characteristics of the coordinate transformation from the physical domain to the computational domain, which is important in our research. Therefore, we apply the simple algebraic transformation proposed above and develop the details of transformation in practical applications in the next sections.

3.1.2 Simplified monotonic transformation

Generally, in a large spatial region, we have many observation points within the region and we know their physical location/coordinates. But we may or may not know the coordinates values of the boundary, which is a non-regular shape and will include all points in its curve. We can use a quadrangle to approximate the boundary of the observation region and then transform the points within the quadrangle into points in a rectangle or square.

Figure 3.2 Simple coordinate transformation using EMAP data. Lat_t and lon_t correspond to the transformed latitude and longitude. (x,y) is one point in the original data and (x',y') is the location of the transformed point.

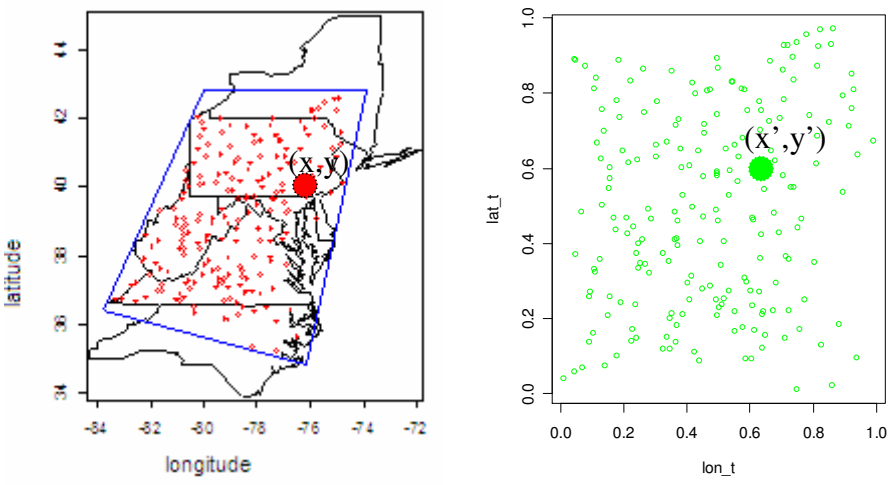


Figure 3.2 shows how the shape of the region is changed after the coordinate transformation. All points in the original physical region have new transformed coordinate values in the new square shaped computational region. When we use a quadrangle to approximate the boundary of an observation region, it is very easy for us

to obtain the linear function for the four boundary lines. Let $(a_l, b_l), (a_r, b_r), (a_t, b_t), (a_b, b_b)$ be the known intercept and slope pairs for left, right, top and bottom boundary lines, which form a quadrangle that makes the observation points within spread over the space as much as possible, and let x_l, x_r, y_t, y_b be the coordinates values of left, right, top and bottom boundary line corresponding to any point (x, y) in the physical domain. Then we obtain

$$x_l = \frac{y - b_l}{a_l},\ x_r = \frac{y - b_r}{a_r},\ y_t = a_t + b_t x \text{ and } y_b = a_b + b_b x. \quad (1)$$

The corresponding coordinate (x', y') in the computational domain can be obtained as $x' = \frac{x - x_l}{x_r - x_l}, y' = \frac{y - y_b}{y_t - y_b}$ (2)

The new variables are linear monotonic transformations corresponding to the original ones. After we get the transformed spatial variables in the computational domain, we will use them as the splitting variables in BCART modeling. After the clustering is done and the terminal node models found, the variables are transformed back to their physical domain. Thus, we can get clustered regions in the physical domain that are more practically effective than the rectangular ones. We call this method of BCART with coordinate transformation of spatial variables an adjusted BCART. In the following section, the method will be applied analyze brook trout data from the eastern United States.

3.2 Application to brook trout dataset

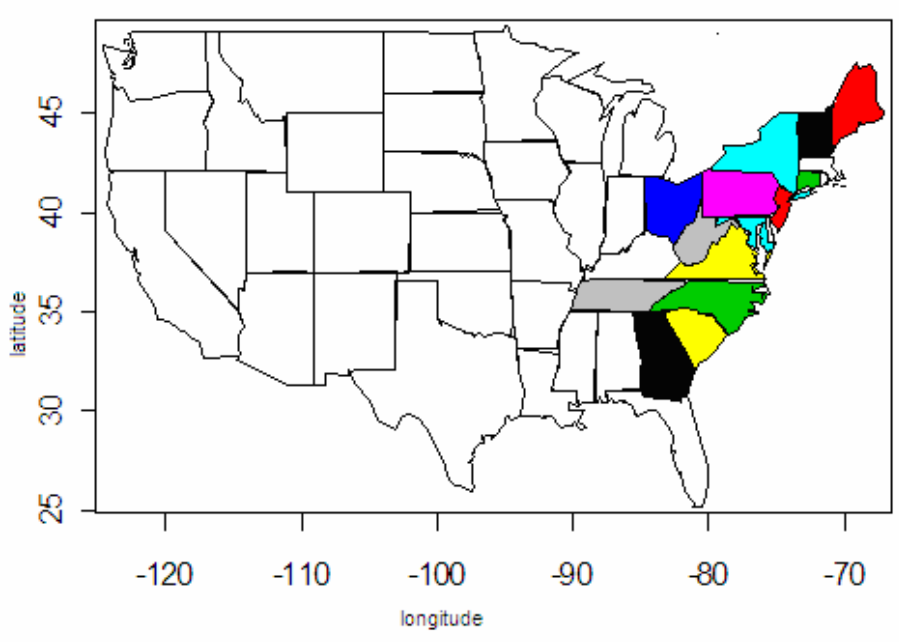
The distribution of brook trout in the eastern United States ranges spatially from Georgia to Maine which covers 16 states. The population of brook trout in this area has declined over the past 200 years because of human perturbations. In many streams, the extirpation (loss) of this species is a serious problem. From the management point of view, it is very important to investigate the possible cause of the population loss and take corresponding action to prevent or restore its loss. Also understanding the relationship

between brook trout population status and the perturbation is essential in developing useful management strategies. Many statistical methods have been applied to model the influential effect on population loss since the Eastern Brook Trout Joint Venture (EBTJV) was formed. This agency was formed to develop a large scale conservation strategy to protect and restore brook trout population and habitat in the eastern region. Hudy *et al.* (2006) laid the ground work for those data analyses. Thieling (2006) investigated several predictive models such as multivariate logistic regression, discriminant analysis and classification trees. Zhang *et al.* (2008) analyzed the same data using the Voronoi diagram-based partition method mentioned in chapter 2 and got an optimal six cluster logistic regression model over the entire region. Zhang's result indicates fairly strong stressor-response relationships within each region and the model is much better than a single regression approach. The modeling result will be compared with my result using an adjusted BCART modeling approach in a later section.

3.2.1 Brook trout data

The dataset addressed in Zhang *et al.* (2008) and Thieling (2006) has a total of 3337 watersheds and 63 candidate stressor metrics. The response variable is the status of brook trout population: extirpation (E) or present (P). The study area is shown in Figure 3.3. Since brook trout in New Hampshire, Vermont, and Maine are uniformly present, the observations within those states are excluded in the study, resulting in 2789 observations, of which 1717 are present and 1072 are extirpated. Zhang *et al*. (2008) selected 4 predictors: Elevation, Road density, Agriculture and Total forest to use in the final model. Since we want to compare our model with Zhang's, we will use the same 4 predictors out of 63 to build our model. For how the original variables are obtained, refer to Thieling (2006) and Zhang *et al*. (2008). These variables were Box-Cox transformed then standardized using the mean and standard deviation of the present group.

Figure 3.3 Study area of brook trout data.



3.2.2 Adjusted BCART modeling

The distribution of the 2789 observations over the region is shown in Figure 3.4a in the physical domain. The region is covered by the observations. The boundary for this region is non-regular. The quadrangle in red is the approximate boundary for this region. The functions for those four boundary lines used are:

$$y_l = 207.25+2.031x_l,\ \ y_r = 19.05+0.875x_r,\ \ y_t = 74.310+0.39x_t,\ \ y_b = 79.652+0.544x_b$$

Based on those four linear functions, the coordinate transformation for each observation within the physical domain can be obtained using formulas (1) and (2). Figure 3.4b shows the computational domain for all observations.

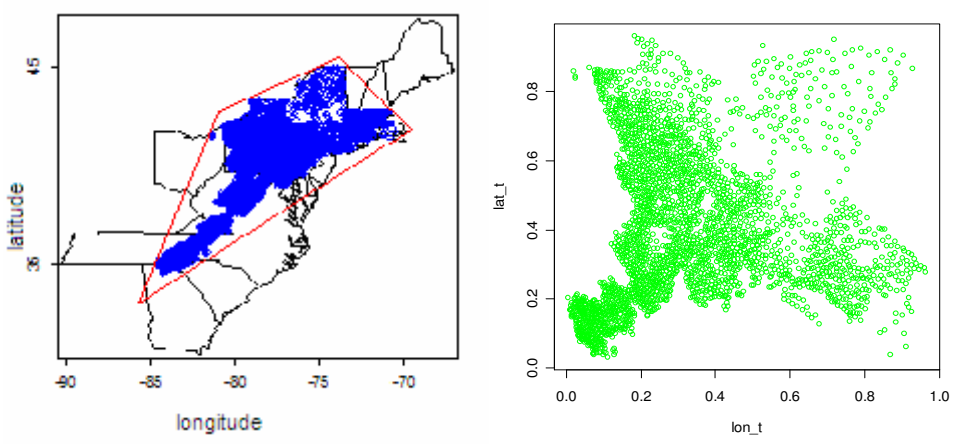
Chapter 2 provided a brief introduction of the Bayesian CART modeling approach. A detailed description of the approach is found in Chipman *et al.* (1998). Here

we sketch the approach again, in the context of our problem. In the BCART approach, we require (i) a formal definition of the parameter space (ii) a choice of prior and (iii) an approach for computing the posterior. The parameter is denoted by the pair $(\mathbf{T}, \boldsymbol{\Theta})$, where $\mathbf{T}$ denotes the tree and $\boldsymbol{\Theta}$ denotes the coefficients in the terminal node logistic regression. We use the most commonly used logistic regression model for categorical (binary) response data as the terminal nodes model

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = x\boldsymbol{\beta} = \beta_0 + \beta_1 x_{elev} + \beta_2 x_{rd} + \beta_3 x_{fr} + \beta_4 x_{ag}$$

where $x_{elev}, x_{ag}, x_{rd}, x_{fr}$ stand for the four selected stressors: Elevation, Agriculture, Road density and Total forest. $\boldsymbol{\beta}$ is a vector of 5 regression parameters with the first parameter corresponding to the intercept term. p is the predicted probability of being extirpated. $\mathbf{T}$ captures the parent-child structure of the tree nodes and the partitioning rule associated with each interior node. The tendency of the Bayesian tree to grow can be defined by the equation $P(\text{node split}|\text{depth}=d)=\alpha(1+\text{d})^{-\beta}$, where α is the base probability of tree growth from splitting a node and β determines the rate at which propensity to split decreases with increased tree size (Chipman *et al.,* 1998). We want to find an optimal 6 or 7 node tree in the computational domain so that the comparison with the result of the Voronoi-diagram based clustering by Zhang *et al.* (2008) can be made. After trying several values of priors, we constrained our range of α to be 0.65-0.85 with step 0.05 and β to be 1-2 with step 0.5. Trees with these prior values give us the number of clusters we need with high posterior probability.

Figure 3.4 Coordinate transformation for brook trout dataset. (a) map of data before transformation, (b) after transformation



(a) (b)

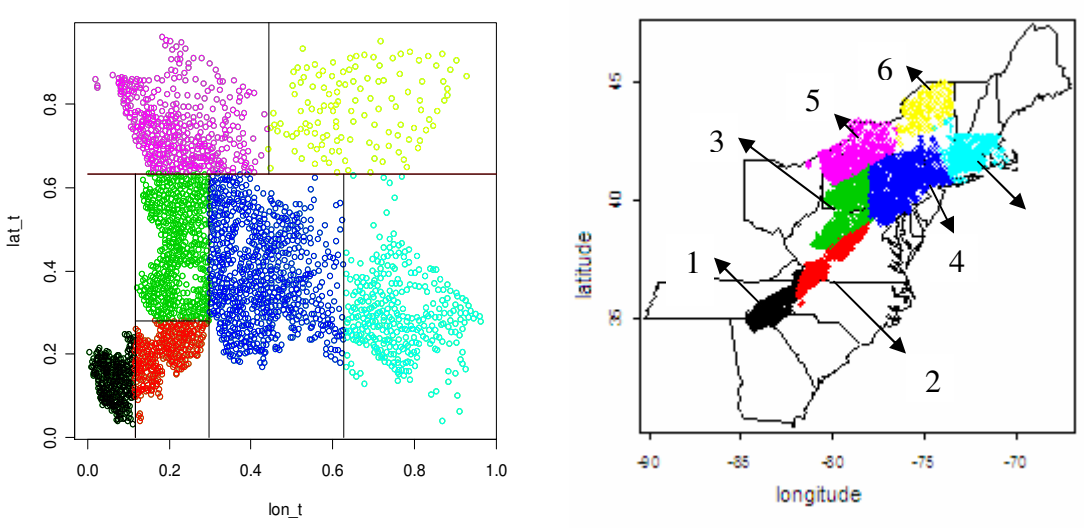
Given $\mathbf{T}$, let k denote the number of terminal nodes. Then $\mathbf{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots \boldsymbol{\beta}_k)$, where $\boldsymbol{\beta}_i$ is the set of coefficients for the logit model fit to the subset of observations corresponding to the ith terminal node. The prior has the form $p(\mathbf{T}, \mathbf{\Theta}) = p(\mathbf{T})\, p(\mathbf{\Theta} \mid \mathbf{T}) = p(\mathbf{T}) \prod_{i=1}^{k} \prod p(\boldsymbol{\beta}_i)$. In the application, we assume equal variance for each terminal node and standardize the explanatory variables first so that the same prior $p(\boldsymbol{\beta})$ can be used independently in each terminal node. This prior is chosen to be a multivariate normal with zero mean and variance matrix proportional to the identity matrix. Hence, all we need to choose is a single prior standard deviation for each of the coefficients in each of the terminal nodes. Chipman *et al.* (2000) provided the details on how to estimate this quantity. The estimate for this quantity is around 2 in our brook trout dataset. We tried prior values equal to 1.5 and 2.

The MCMC simulation was started at the tree with just the root node and run for 8000 iterations. The chain was started 3 different times (8000 iterations each time). We selected the tree with the highest integrated likelihood amongst all those visited.

3.2.3 7-cluster partition model with adjusted BCART and BCART

After running the adjusted BCART for the prior combinations, a tree with a 7 node clustering in the computational domain resulted in the highest posterior likelihood and was selected. It is shown in Figure 3.5a. The tree prior uses parameters $\alpha = 0.75, \beta = 2$ and the model parameter prior is $\hat{\sigma}_\beta = 1.5$. We can see that the region is clustered into 7 rectangular regions in the computational domain. The clustering solution then is transformed back to the physical domain. Figure 3.5b is the clustered region in the physical domain. The corresponding terminal node models and the model without classification are illustrated in Tables 3.1 and 3.2.

Figure 3.5 Graphical display of adjusted BCART 7-cluster modeling result. (a) clusters in the transformed coordinate space. (b) clusters in the original space.



(a) (b)

Table 3.1 Parameter estimates for logistic models with 7-clusters partitioned by adjusted BCART. Refer to Figure 3.5 for locations of clusters.

Parameter estimates for logistic model on the 7-clusters						
Cluster	Present/Absent	Intercept	Elevation	Road Dens	Ag	Forest
1	155/177	3.266*	−1.983*	0.348	0.003	−0.461
2	191/123	0.958^	−2.271*	−0.069	0.116	−0.520
3	326/178	−1.365*	−3.098*	0.668*	0.942^	0.312
4	433/344	−2.688*	−1.602*	0.269	0.372^	−0.477*
5	210/200	−3.147*	−3.371*	0.997^	1.741*	0.422
6	126/0	--	--	--	--	--
7	276/50	−4.007*	−0.729	0.660	−0.087	−0.868

^ significance level $\leq$ 0.05

* significance level $\leq$ 0.005

Table 3.2 Parameter estimates for logistic model without partition.

Parameter estimates for logistic model on the entire dataset						
N	Present/Absent	Intercept	Elevation	Road dens	Ag	Forest
2789	1717/1072	−1.016*	0.214*	0.448*	0.487*	−0.349*

Figure 3.6 is the 7-cluster BCART solution without spatial transformation. Table 3.3 gives the parameter estimates for the logistic model using the 7-clusters BCART result. As we can see, the clusters consisted of 7 rectangles. Some clusters have "fingers" which are not appropriate, e.g. the light blue region covers most area of Massachusetts and Connecticut, but it has an additional narrow area that is along the eastern border of New York which does not seem to be a reasonable partitioning of the region.

It is commonly known that brook trout prefer cooler temperatures (Hudy, personal communication). The result of the benchmark model (modeling the extirpation) is misleading because the estimate of the coefficient of elevation has a positive sign. This model has a result that is contradictory with existing knowledge. The result using BCART and the adjusted BCART is much more reasonable than the previous result in that all the covariates of elevation for different clusters have parameter estimates that are negative. This tells us that that the models within each subregion are better in the sense that they are not inconsistent with existing knowledge. But, how do we know which approach is better than the other statistically? We need a model performance measure to make the comparison among different partition approaches. In the next section, we will

introduce a model performance measure and compare the results by different partition approaches.

Figure 3.6 BCART 7-cluster modeling result without spatial transformation.

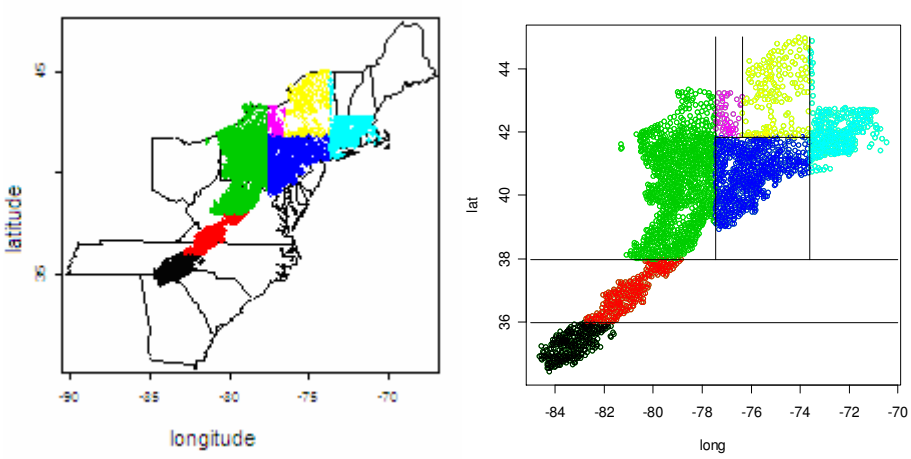


Table 3.3 Parameter estimates for logistic model on the 7-clusters using BCART.

Parameter estimates for logistic model on the 7-clusters using BCART						
cluster	Present/Absent	Intercept	Elevation	Road Dens	Ag	Forest
1	123/164	3.694*	-1.784*	0.170	0.373	-0.784
2	163/96	0.963^	-1.595*	-0.283	0.059	-0.866
3	653/420	-1.848*	-2.269*	0.942*	1.238^	0.384
4	332/298	-2.701*	-1.601*	0.179	0.226^	-0.619*
5	10/40	-6.696*	-2.778	3.904	3.073	5.819
6	161/4	-3.963*	-0.589*	0.504*	-0.201*	-0.443
7	275/50	-1.204	-2.676	-0.023	2.136	1.695

^ significance level $\leq$ 0.05

* significance level $\leq$ 0.005

3.3 Model performance and comparison

3.3.1 Performance assessment using ROC curve

A confusion matrix can be used to evaluate the performance of logistic regression for classification. Table 3.4 is one example for a 2-class problem. The columns are the predicted class and the rows are the actual class. True Negatives (TN) is the number of negative examples correctly classified, False Positives *(FP)* is the number of negative examples incorrectly classified as positive, False Negatives (*FN)* is the number of

positive examples incorrectly classified as negative and True Positives *(TP)* is the number of positive examples correctly classified. In logistic regression, we use predictive accuracy as the general performance measure, which has the form *Accuracy = (TP+TN)/(TP+FP+TN+FN).* In the context of balanced datasets and equal error costs, it is reasonable to use this as a performance metric. In the presence of unbalanced datasets with unequal error costs, which is our situation, it is more appropriate to use the Receiver Operating Characteristic (ROC) curve or other similar techniques (Fawcett, 2003).

The ROC curve is a graphic display that indicates the predictive accuracy of a classification model. ROC graphs are two-dimensional graphs in which the *TP* rate is plotted on the Y axis and *FP* rate is plotted on the X axis. The ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). The Area Under the Curve (AUC) is an accepted traditional performance metric for a ROC curve (Pepe *et al.*, 2005). Generally, the larger the AUC, the better the classification model is. For our problem, we use the estimator for this area $\text{AUC} = \frac{1}{n_E n_P} \sum_{i \in E, j \in P} I(p_i > p_j)$, 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. For the adjusted BCART model, we use the weighted average AUC measurement used by Zhang to assess the model performance.

$$\text{AUC} = \frac{\sum_m AUC_m n_m + \sum_l n_l}{N} \tag{3}$$

Here l indexes the cluster which has more than 90% of the observations belonging to one category. We call this cluster an overwhelming cluster. m indexes all other clusters. N is the total sample size and n_l, n_m are the sample sizes for the clusters indexed by l or m.

Table 3.4 Confusion Matrix for two class problem.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

3.3.2 Comparison with BCART and benchmark model

From Table 3.1, we can see that cluster 7 has no brook trout sites that are extirpated and cluster 6 only has a small amount of extirpated sites. We can combine these two clusters as an un-extirpated region to form an overwhelming cluster, called region 67. We use 10-fold validation to calculate the AUC for the previous 7-node solution, the combined 6-node solution and 1–cluster solution for the entire region. Table 3.5 is the performance for the single logistic model. Table 3.6 is the model performance for the 7-cluster and 6-cluster adjusted BCART model. AUC^* is the quantity when a proportion of 0.90 is used in defining an overwhelming cluster. We can see that the BCART clustering model has a much better AUC or AUC^* within each region and the total average AUC increases from 0.77 for 1-cluster benchmark model to 0.90 for the 6-node partition solution, a 13% increasing of correct prediction. Table 3.7 summarizes the10-fold crossvalidation results for the final 7-cluster BCART solution and the 6-cluster BCART solution. Both the adjusted BCART and the BCART approaches have good performance in terms of the AUC and are much better than the benchmark model.

Table 3.5 Ten-fold crossvalidation results for the benchmark model.

Single model approach		
No. obs	AUC	AUC*
2789	**0.7727**	**0.7727**

Table 3.6 Ten-fold crossvalidation results for the final 7-cluster adjusted BCART solution (a) and 6-cluster adjusted BCART solution (b). AUC* is the quantity when a proportion of 0.90 is used in defining an overwhelming cluster.

7-node adjusted BCART model			
Cluster	No.obs	AUC	AUC* *
1	332	0.8353	0.8353
2	314	0.8125	0.8125
3	504	0.9136	0.9136
4	777	0.8773	0.8773
5	410	0.9188	0.9188
6	126	1.0000	0.9000
7	326	0.7684	0.7684
Total	2789	**0.8705**	**0.8660**

(a)

6-nodes adjusted BCART model			
Cluster	No.obs	AUC	AUC*
1	332	0.8353	0.8353
2	314	0.8125	0.8125
3	504	0.9136	0.9136
4	777	0.8773	0.8773
5	410	0.9188	0.9188
67	452	1.0000	0.9000
total	2789	**0.8988**	**0.8826**

(b)

Table 3.7 Ten-fold crossvalidation results for the final 7-cluster BCART solution and 6-cluster BCART solution. AUC* is the quantity when a proportion of 0.90 is used in defining an overwhelming cluster.

7-nodes BCART model			
Cluster	No.obs	AUC	AUC*
1	287	0.8305	0.8305
2	259	0.7753	0.7753
3	1073	0.9141	0.9141
4	630	0.8713	0.8713
5	50	0.4075	0.4075
6	165	1.0000	0.9000
7	325	0.7871	0.7871
total	2789	**0.8598**	**0.8502**

(a)

6-nodes BCART model			
Cluster	No.obs	AUC	AUC *
1	287	0.8305	0.8305
2	259	0.7753	0.7753
3	1073	0.9141	0.9141
4	630	0.8713	0.8713
5	50	0.4050	0.4050
67	490	0.8222	0.8222
total	2789	**0.8583**	**0.8583**

(b)

3.3.3 Comparison with Voronoi diagram-based partition model

Zhang *et al.* (2008) analyzed this brook trout data using a Voronoi-diagram based partition method introduced in Chapter 2. The criterion for model selection and assessment is the AUC in formula (3) of the ten fold crossvalidation. The final 6-cluster partition model was selected after 20,000 Monte Carlo simulations. Table 3.9 provides the model performance evaluation and Table 3.10 describes the models within each cluster. The performance of the 6-cluster partition based on the Voronoi diagram technique is as good as that of the 6-cluster partition using the adjusted BCART approach. We can see that both modeling approaches greatly improve model performance relative to the benchmark model. Table 3.11 tells us that both the adjusted BCART and the Voronoi-diagram based partition approaches have good performance in terms of AUC and are much better than the benchmark model. Adjusted BCART has an AUC of value 0.90 while Voronoi diagram based partition has an AUC of 0.89. The performance in prediction accuracy ability is very close.

Table 3.8 Model performance evaluation for the 6-cluster Voronoi diagram-based partition.

6-cluster Voronoi diagram-based partition			
Cluster	No.obs	AUC	AUC *
1	403	0.81	0.81
2	314	0.87	0.87
3	633	0.91	0.91
4	757	0.85	0.85
5	251	0.93	0.93
6	431	1.00	0.90
total	2789	**0.89**	**0.88**

Figure 3.7 shows the clustering comparison between Voronoi-diagram modeling and adjusted BCART modeling on the map. We can see that the partitions from both approaches are geographically similar.

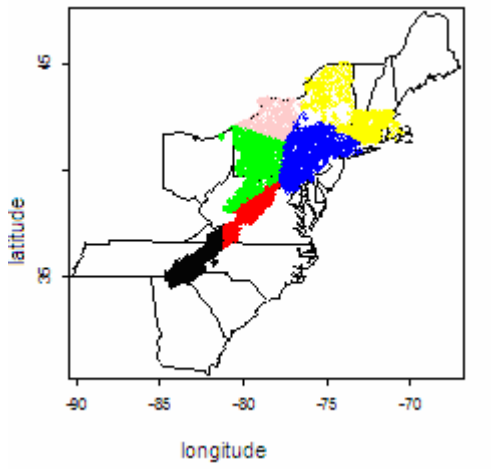
Table 3.9 Parameter estimates for the 6-cluster partition based on Voronoi technique.

Parameter estimates for logistic models based on 6-cluster solution						
cluster	present/extirpated	Intercept	Elevation	Road dens	Ag	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

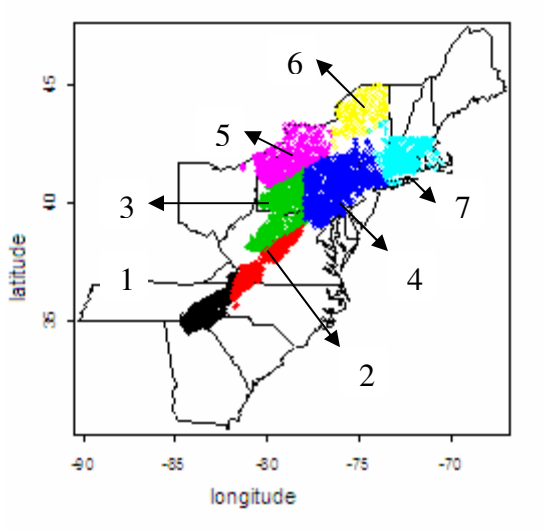
^ significance level $\leq$ 0.05

* significance level $\leq$ 0.005

Figure 3.7 Comparison of partitions between (a) the Voronoi-diagram modeling and (b) the adjusted BCART modeling in map.



(a)



(b)

Table 3.10 Comparison of model performance (AUC) for the 6-cluster partitions based on ABCART, Voronoi-diagram partitioning and the benchmark model.

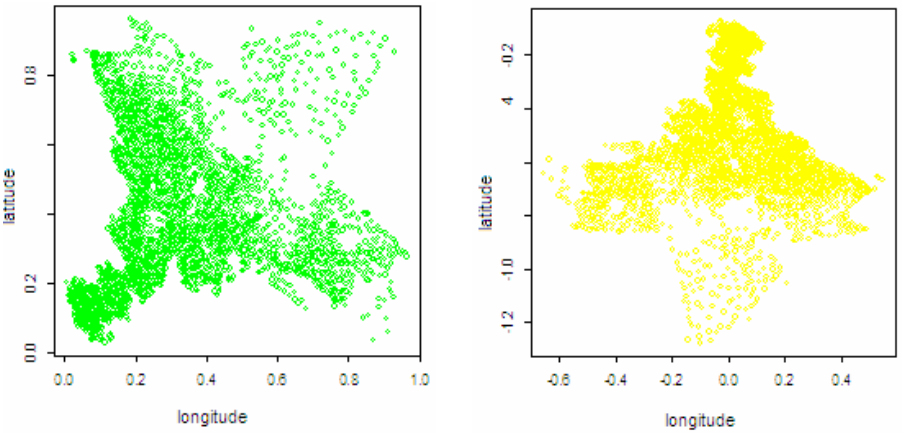
Model	ABCART	Spatial Partition	Benchmark
AUC	0.90	0.89	0.77

3.4 Model validation and conclusion

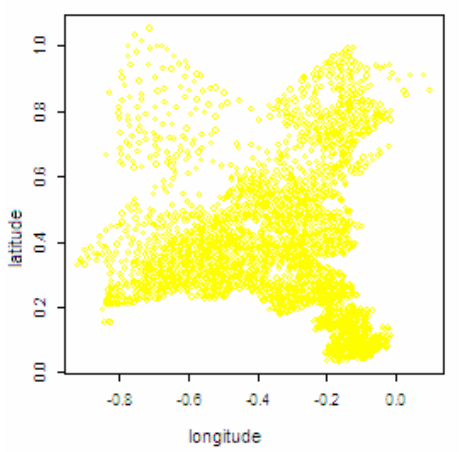
The adjusted BCART model has the best performance in terms of AUC values since the transformed observations are more evenly and fully distributed on the splitting space. To verify that the splitting in the adjusted BCART model is the best, we rotate the transformed observations through several angles to illustrate that the partition on a more evenly distributed space will have a better performance. Let the rotation angle be α=15, 30, 45, 60, 75 and 90. We then compare the average AUC values for the partitions from each different angle. The six or seven 'best' clusters with highest posterior probability were chosen for each rotation angle. After observation (x, y) has been rotated through an angle α, its new coordinates are (x', y'). Let $\mathbf{(x, y)}$ be the coordinates of all observations in a computational region, then, we have the new coordinates $\begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$.

Figure 3.8 shows the distribution of the observations for brook trout data on the computational space after rotation through an angle α. We can see that, without rotation (α=0), the ratio of observation area over the rectangular area is the largest. This means that the observations are the most evenly distributed over the splitting space in this case. Table 3.11 gives the comparison of the AUC values for each different angle. We can see that α=0 has the highest AUC. The details on clusters for each angle and the models for each cluster in that angle are presented in the appendix.

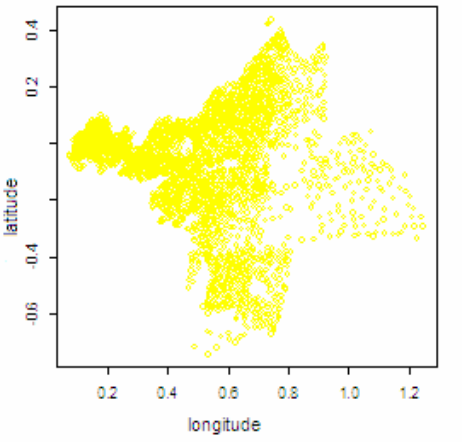
Figure 3.8 The distributions of the observations on the computational space after rotation through an angle α. α=0, 15, 30, 45, 60, 75 and 90.



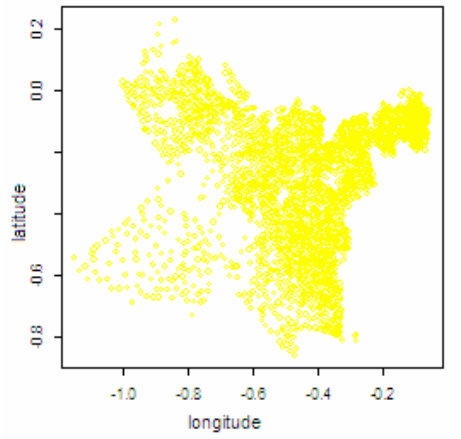
(a) α=0 (b) α=15



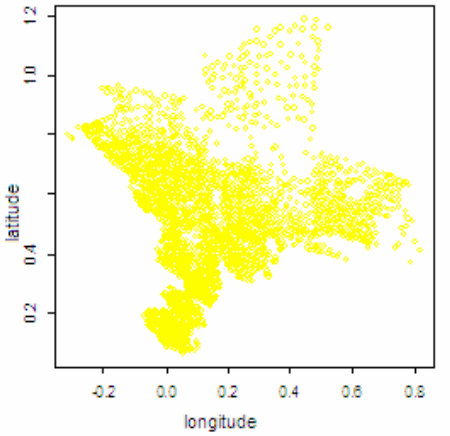
(c) α =30



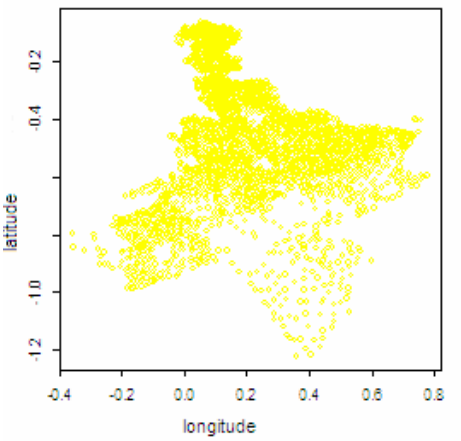
(d) α =45



(e) α =60



(f) α =75



(g) α =90

Table 3.11 AUC* values after rotation for each angle.

Angle	Best partition	AUC*
0	6	0.8826
15	7	0.8573
30	7	0.8619
45	7	0.8676
60	7	0.8689
75	6	0.8632
90	7	0.8572

3.4.1 Concluding remarks

An adjusted BCART approach can be used as an alternative partition modeling approach to the Voronoi diagram based approach. It has better results than the benchmark model and is as good as the Voronoi diagram approach. To summarize, the adjusted BCART has the following nice properties

(1) The partition by adjusted BCART has better performance characteristics. From the Appendix we can see that all the partitions with rotation through different angles have models that are better than the benchmark model. All covariates for elevation are negative and the AUC*'s are around 0.86.

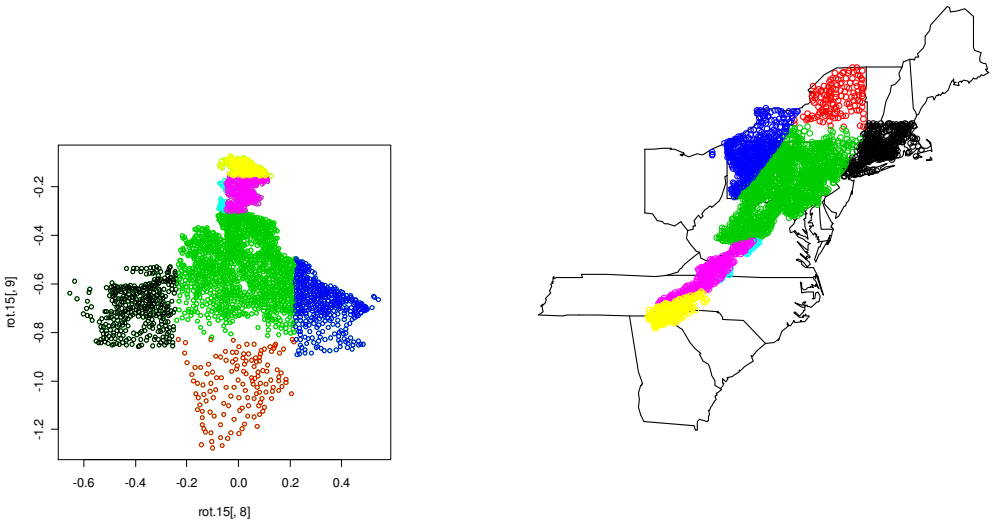
(2) The performance in terms of the AUC by adjusted BCART is as good as the Voronoi-diagram based partition approach.

(3) The performance of the adjusted BCART is better than that of the BCART for the analysis in terms of AUC. The adjusted BCART approach helps to stretch the spatial distribution of data more evenly over the space and thus partition the space into subregions more effectively. This claim was verified by the rotation transformations in the previous section. The AUC* for the zero degree rotation is the highest. Also rotation avoids the 'finger area' for some clusters that was seen to occur in the BCART partition.

The down side of this adjusted BCART approach is the lack of flexibility to be extended to multivariate analysis or any other analysis that requires an exact distribution of model within terminal nodes. A new approach must be explored for the multivariate situation, which is one of the important research interests in ecological study.

Appendix

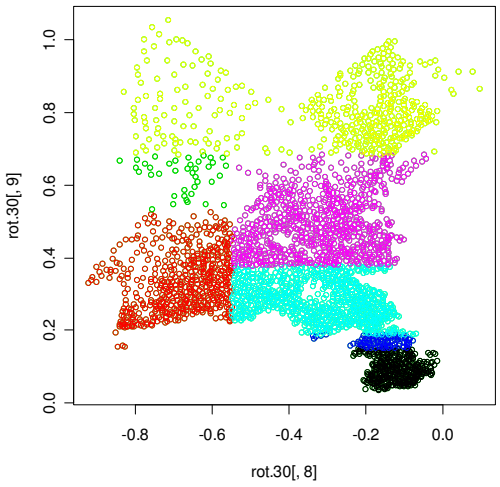
Figure A3.1: Optimal partitions and models for rotation transformation with different degrees

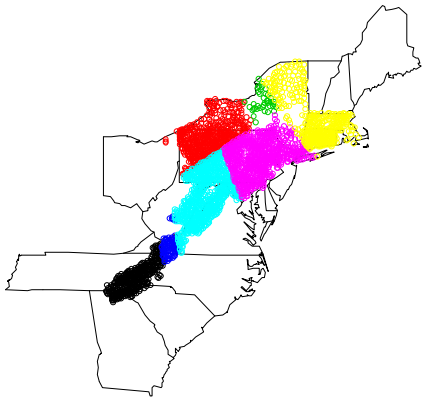


Parameter estimates for logistic models based on 7-cluster solution							
Partition	Present/ Absent	Int	Elev	Road density	Ag	Total forest	AUC*
361 Black	277/84	-3.9024	-0.3909	0.8179	-0.0819	-0.3801	0.8500
138 Red	136/2	--	--	--	--	--	0.9000
1340 Green	849/491	-1.7489	-0.9494	0.3133	0.6108	-0.2319	0.8578
412 Blue	176/236	-3.9347	-4.6461	1.3761	1.9317	0.6042	0.9401
27 Lake	26/1	--	--	--	--	--	0.9000
290 Pink	179/125	1.4438	-1.2478	0.0291	-0.4568	-0.9011	0.7375
221 Yellow	88/133	3.8122	-2.0813	0.3018	0.1521	-0.9414	0.8388

AUC=2403.9/2789=0.8573

30 degrees

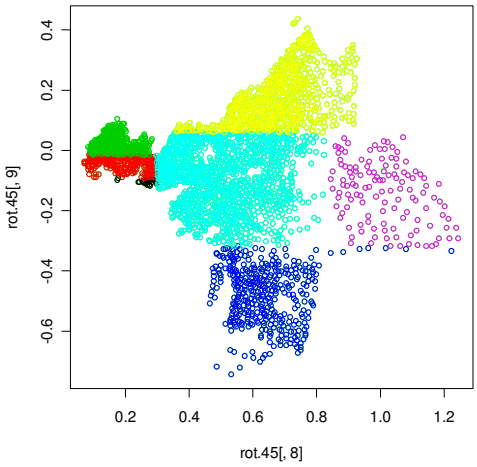


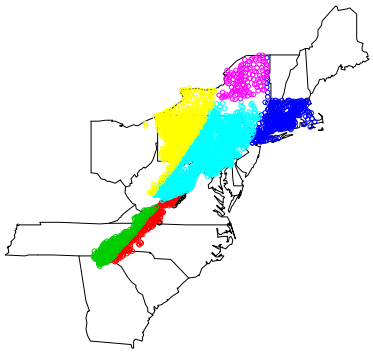


Parameter estimates for logistic models based on 7-cluster solution							
Partition	Present/ Absent	Int	Elev	Road density	Ag	Total forest	AUC*
361 Black	178/183	3.1535	-1.9854	0.3069	-0.2780	-0.7832	0.8367
496 Red	253/243	-3.4962	-4.1455	1.3566	1.9792	0.7297	0.9359
39 Green	38/1	--	--	--	--	--	0.9000
89 Blue	51/38	0.1718	-0.2343	-0.3898	-0.7762	-1.5296	0.5609
690 Lake	430/260	-0.7629	-1.7954	0.5864	0.6599	0.2177	0.8666
710 Pink	406/304	-2.6542	-1.2719	0.3015	0.4559	-0.5075	0.8749
404 Yellow	361/43	-4.8157	-1.2012	0.6406	0.0098	-0.3513	0.8256

AUC=2403.9/2789=0.8619

45 degrees



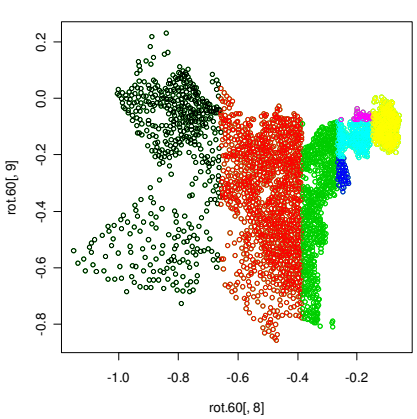


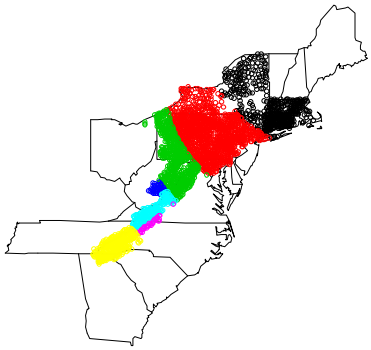
Parameter estimates for logistic models based on 7-cluster solution

Cluster	Present /Absent	Int	Elev	Road density	Ag	Total forest	AUC*
20 Black	20/0	--	--	--	--	--	0.9000
215 Red	116/99	2.2895	-2.6455	0.2305	-1.2982	-1.5575	0.7861
301 Green	142/159	3.9512	-2.3707	0.5453	-0.3557	-0.7389	0.8445
376 Blue	290/86	-4.0301	-0.5043	0.8846	-0.0356	-0.3137	0.8443
1132 Lake	692/440	-1.6904	-0.9307	0.2339	0.4428	-0.3571	0.8575
125 Pink	125/0	--	--	--	--	--	0.9000
620 Yellow	332/288	-2.6931	-3.0516	1.2362	1.6781	0.4107	0.9343

AUC=2419.88/2789=0.8676

60 degrees

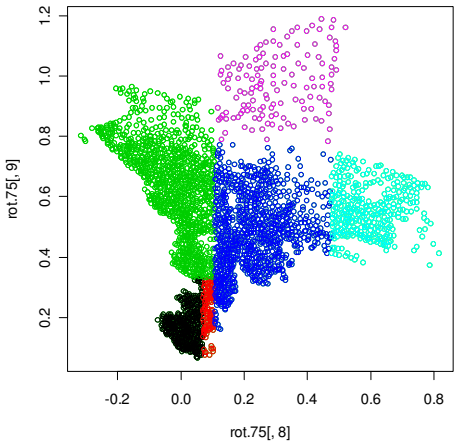


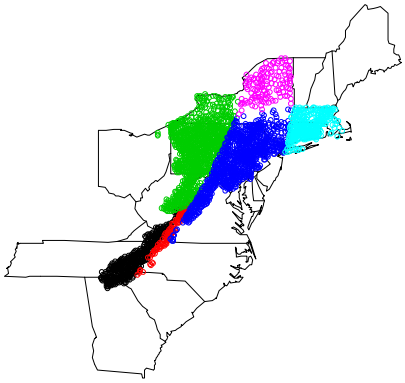


Parameter estimates for logistic models based on 7-cluster solution							
Cluster	Present /Absent	Int	Elev	Road density	Ag	Total forest	AUC*
497 Black	439/58	-4.6850	-1.0157	1.003	-0.0494	-0.0653	0.8474
1150 Red	671/479	-2.0588	-1.1731	0.3353	0.5512	-0.3009	0.8623
568 green	291/277	-0.6934	-2.4593	0.3309	-0.1111	-1.2305	0.9401
43 Blue	41/2	--	--	--	--	--	0.9000
190 Lake	104/86	1.3231	-1.7892	-0.1069	0.3777	-0.3821	0.7795
33 Pink	29/4	--	--	--	--	--	0.9000
308 Yellow	142/166	3.5362	-1.9417	0.2744	0.1703	-0.6991	0.8455

AUC= 2423/2789=0.8689

75 degrees

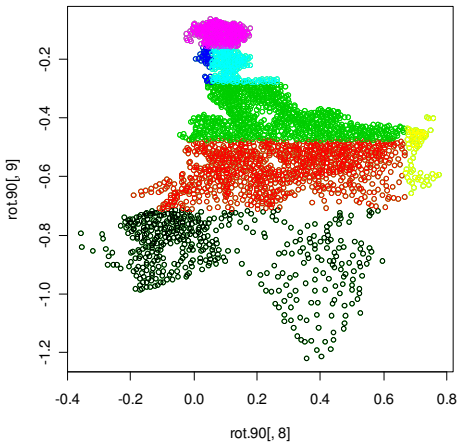


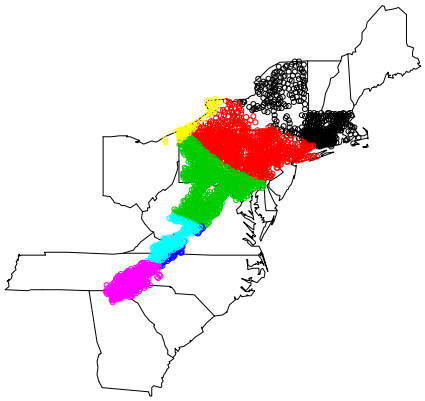


Parameter estimates for logistic models based on 6-cluster solution							
Cluster	Present /Absent	Int	Elev	Road density	Ag	Total forest	AUC*
413 Black	203/210	2.9145	-2.0296	0.4781	-0.3779	-0.5386	0.8202
108 Red	69/39	1.7976	-3.6644	-0.1984	0.0478	-0.3009	0.8623
935 Green	592/343	-2.0546	-2.5933	0.8209	1.1436	0.2430	0.9133
884 Blue	459/425	-1.7946	-0.6195	0.3969	0.4911	-0.4551	0.8456
315 Lake	260/55	-4.1673	-0.6715	0.5970	-0.2149	-0.4162	0.8052
134 Pink	134/0	--	--	--	--	--	0.9000

AUC= 2407.55/2789=0.8632

90 degrees





Parameter estimates for logistic models based on 7-cluster solution

cluster	Present /Absent	Int	Elev	Road density	Ag	Total forest	AUC*
451 Black	408/43	-5.414	-1.8538	0.4989	0.1488	-0.3798	0.8363
918 Red	614/304	-2.3437	-0.9249	0.5033	0.6883	-0.3036	0.8692
807 Green	402/405	-1.0346	-1.7694	0.5057	0.6943	0.1748	0.8856
41 Blue	36/5	1.3534	-5.7280	1.5718	-1.2895	1.0328	0.5305
196 Lake	110/86	1.0656	-1.5384	-0.0327	0.3910	-0.5227	0.7809
307 Pink	142/165	3.4505	-1.8959	0.2972	0.1509	-0.6688	0.8436
69 Yellow	5/64	--	--	--	--	--	0.9000

AUC=2385/2789=0.8553

Table A3.1 Brook trout distribution for binary response within each state.

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

4 Spatial Partition Modeling with Multicategorical Responses

4.1 Introduction

In ecological studies, we sometimes have observations with multinomial categorical responses. For example, in water quality assessment, the response variable could be watershed impairment status with values heavily impaired, impaired, lightly impaired and intact. The relationship between the multinomial categorical response and potential stressors can be evaluated through multinomial logit/probit models. The simplest multinomial response is the binary response which is typically modeled using logistic regression. For the situation where the response category has more than two levels, we can collapse many levels into two so that the logistic regression theory and methods can be used. Zhang *et al.* (2008) proposed a partition modeling approach for logistic regression by collapsing the multicategory responses into a binary response. The binary response impairment/non-impairment or present/absent (for species) is a very coarse classification for the water quality status in some situations. Collapsing categories may sometimes detract from the main interest of the investigation (Lawal, 2003). Often we are more interested in knowing the specific watershed or species status within clusters so that the corresponding suggestions can be proposed by the watershed managers. In this chapter we propose a multinomial category partition modeling approach by extending the binary logistic model to a multinomial logit model and then apply it to the brook trout data. Agresti (2002) described the logistic regression and multinomial regression in detail in his book. We give a brief introduction on these two modeling approaches at the beginning of this chapter. Then we use an Average Fraction Correctly Classified for Fit (AFCCF) criterion (Olsen, 2003) to select the partition model and introduce the partition modeling process for the multinomial logit model. The approach is finally applied to the brook trout dataset.

4.2 Method

4.2.1 Logistic regression model

Many response variables are binary and represent present/absent or success/failure by 1 and 0, respectively. The binary variable y is assumed to have a Bernoulli distribution. The probability of success is $P(y=1)=p$ and the probability of failure is $P(y=0)=1-p$. Consequently the probability mass function is $f(y,p)=p^{y}(1-p)^{1-y}=(1-p)\exp(y\ln(\frac{p}{1-p}))$. From a generalized linear model (GLM) point of view, the natural parameter, $\ln(\frac{p}{1-p})$, is the log odds of the response and is called the logit of p. In a GLM, we want to model the probability of success as $g(p)=\mathbf{x\beta}$, where $\mathbf{x}$ is the vector of stressor variables, $\boldsymbol{\beta}$ is a vector of parameters and g is a link function which describes the relationship between the linear predictor $\mathbf{x\beta}$ and the expected values of the response variable. If the link function used is the logit, we will call the model a logistic model, i.e. $\ln(\frac{p}{1-p})=\alpha+\mathbf{x\beta}$. After exponentiating both sides of this equation, we find that the odds of success increases multiplicatively by e^{β} for every 1 unit increase in $\mathbf{x}$ (Lawal, 2003).

4.2.2 Multicategory logit models

In the logistic model as described in the last section, we restricted the response variable to be dichotomous. Now let us consider a response variable, y, with J levels. Let $p_j(\mathbf{x})=P(y=j\mid\mathbf{x})$ for explanatory variables $\mathbf{x}$ with $\sum_j p_j(\mathbf{x})=1$. $\mathbf{x}$ can be quantitative, qualitative, or both.

When $J=2$, there is only one logit we can form and at each combination of the explanatory variable we assume that the data comes from a binomial distribution. When $J>2$, there are $J(J-1)/2$ logits that we can form, but only $J-1$ of them are

non-redundant. There are different ways to form the non-redundant logits, each of which results in a "dichotomizing" of the response variable. The way to form the logit will partly depend on whether y is ordinal or nominal (Agresti, 2002).

4.2.2.1 Baseline Category Logit Model for Nominal Response Variables

The baseline category logit model can be viewed as an extension of the binary logistic regression model. The model gives a simultaneous representation of the odds of being in one category relative to being in another category, for all pairs of categories. In this model, we choose one of the categories as the "baseline". For convenience, the last level (i.e. the J^{th} level) of the response variable is usually used as the baseline (Anderson, 2006).

The baseline category logit model with one explanatory variable, x, can be expressed as:

$$\log\left(\frac{p_j}{p_J}\right) = \alpha_j + \beta_j x$$

for $j = 1, 2...J-1$. For $J = 2$, this model becomes the regular binary logistic regression model. For $J > 2$, α and β can be different depending on which two categories are being compared. The odds for any pair of categories of y are a function of the parameters of the model. Suppose we have a response variable which has 3 categories, then we have 2 non-redundant logits:

$$\log\left(\frac{p_1}{p_3}\right) = \alpha_1 + \beta_1 x$$

$$\log\left(\frac{p_2}{p_3}\right) = \alpha_2 + \beta_2 x$$

The logit for categories 1 and 2 is:

$$\log\left(\frac{p_1}{p_2}\right) = \log\left(\frac{p_1 / p_3}{p_2 / p_3}\right)$$
$$= \log(p_1 / p_3) - \log(p_2 / p_3) = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x$$

We can interpret the parameters of the model in terms of odds ratios, given an increase in x. Just as in binary logistic regression, we can also interpret the parameters of the model in terms of probabilities. The probability of a response being in j^{th} category is

$$p_j = \frac{\exp(\alpha_j + \beta_j x)}{\sum_{j=1}^{J} \exp(\alpha_j + \beta_j x)}$$

For the baseline category, we have $\alpha_J = \beta_J = 0$, which is an identification constraint. Furthermore, the denominator $\sum_{j=1}^{J} \exp(\alpha_j + \beta_j x)$ ensures the sum of probabilities equal to 1. We can use these functions to plot the probabilities versus x. In addition, more explanatory variables can be added to the model and these variables can either be categorical or numeric. The baseline category logit model can work well when the categories of the response variables are ordered, but the proportional odds model can be often be better.

4.2.2.2 Proportional Odds Model for Ordinal Response Variables

According to Anderson (2006), when the response variable is ordered we can use the ordering of the categories in forming the logits. For this model we use cumulative probabilities of a response variable y falls in category j or below. In other words, $P(y \le j) = p_1 + p_2 + ... + p_j,\ j = 1, 2, ..., J$. Cumulative probabilities reflect the ordering of the categories and are used to form cumulative logits. The cumulative logit can be expressed as:

$$\log\left(\frac{P(y \le j)}{P(y > j)}\right) = \log\left(\frac{P(y \le j)}{1 - P(y \le j)}\right) = \log\left(\frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + \ldots + p_J}\right)$$

The response variable collapses into two categories and the proportional odds ratio is:

$$\log\left(\frac{P(y \le j)}{1 - P(y \le j)}\right) = \alpha_j + \boldsymbol{\beta}\mathbf{x}$$

Thus, cumulative probabilities are given by:

$$P(y \le j) = \frac{\exp(\alpha_j + \boldsymbol{\beta}\mathbf{x})}{1 + \exp(\alpha_j + \boldsymbol{\beta}\mathbf{x})}$$

The probability of being in category j can be computed by taking differences between cumulative probabilities. That is

$$P(y = j) = P(y \le j) - P(y \le j - 1) \text{ for } j = 2, ..., J$$

$$P(y = 1) = P(y \le 1)$$

Since the coefficient $\boldsymbol{\beta}$ is assumed to be the same for all logistic equations, this model is called the proportional odds model.

4.2.3 AFCCF model selection criterion

The Average Fraction Correctly Classified for Fit (AFCCF) is a measure of the classification model's ability to make a correct prediction. The AFCCF is similar to the R-square value of the linear model in that higher values indicate a closer match between the model and the data (Olsen, 2003). The formula for logistic regression is

$$AFCCF = \frac{\sum_{i=1}^{n} y_i \hat{p}_i + \sum_{i=1}^{n} (1 - y_i)(1 - \hat{p}_i)}{n}, \quad i = 1, 2, ..., n$$

where y_i is the value of i^{th} observation. For logistic regression, if the i^{th} observation is a success, then $\hat{p}_i$ is the prediction for the probability of success. On the other hand, if the i^{th} observation is a failure, then $1 - \hat{p}_i$ is the prediction for the probability of failure.

$\sum_{i=1}^{n} y_i \hat{p}_i$ is the sum of the estimated probability of success and $\sum_{i=1}^{n} (1-y_i)(1-\hat{p}_i)$ is the sum of the estimated probability of failure. The better the prediction performance is, the higher the AFCCF value. An AFCCF that is closer to 1 for a model reflects better agreement between the model and the data than an AFCCF that is closer to 0.

For the multicategory response model, the formula can be written as

$$AFCCF = \frac{\sum_{i=1}^{n} I(y_i \subset 1)\hat{p}_{i1} + \sum_{i=1}^{n} I(y_i \subset 2)\hat{p}_{i2} + ... \sum_{i=1}^{n} I(y_i \subset J)\hat{p}_{iJ}}{n}$$

where y_i is the value of i^{th} observation. $I(y_i \subset j)$ is an indicator function to judge if the i^{th} observation belongs to j^{th} category. If it is true, then $I(y_i \subset j)$ has value 1. Otherwise, it has value 0. $\sum_{i=1}^{n} I(y_i \subset j)\hat{p}_{ij}$ is the sum of the estimated probability of the j^{th} category.

4.2.4 Spatial partition model selection using AFCCF for multicategory logit model

The Voronoi diagram based partition method proposed in chapter 2 is a method for clustering sites to maximize the strength of cluster-wise stressor-response relationships within clusters. For the current application, we expect the multicategory logit relationships between response and stressors to be different for different regions. The Voronoi diagram is used to randomly assign sites to one of *k* clusters based on certain measurements (ie. latitude and longitude or width and elevation) and the AFCCF criterion using 5-fold crossvalidation is used with model selection to get the optimal partition. Suppose that we have *K* clusters for one Voronoi tessellation (partition), the AFCCF for the k^{th} cluster among the *K*-cluster model can be written as

$$AFCCF_k = \frac{\sum_{i=1}^{n_k} I(y_i \subset 1)\hat{p}_{i1} + \sum_{i=1}^{n_k} I(y_i \subset 2)\hat{p}_{i2} + ... \sum_{i=1}^{n_k} I(y_i \subset J)\hat{p}_{iJ}}{n_k}, \quad k = 1, 2, ..., K$$

Here n_k is the number of observations for the k^{th} cluster. Then the AFCCF for the *K*-cluster partition model is the average of AFCCF for each cluster, $AFCCF = (\sum_{k=1}^{K} AFCCF_k)/K$.

Usually the AFCCF value increases with the number of partitions. In other words, the larger the value of *K,* the higher the AFCCF value. This proportional feature is caused by using the same data to build a model and to obtain predictions of responses. Using 5-fold crossvalidation, we separate the modeling and the prediction data and thus reduce the prediction bias of the AFCCF criterion. The partition which has relative high AFCCF value from 5-fold crossvalidation is the one which has less overall prediction error. When the number of partitions exceeds the underlying number of clusters, the prediction error increases and the AFCCF value of 5-fold crossvalidation decreases

The spatial partition modeling procedure for a multicategory logit model is described as follows:

1. Partition the two-dimensional spatial region of data into *k* non-overlapping clusters using the Voronoi diagram technique.
2. Calculate the AFCCF value from 5-fold crossvalidation for this *k*-cluster model.
3. Repeat steps 1 and 2 enough times such that the optimal *k*-cluster partition is found. The optimal *k*-cluster partition is the partition which has the highest AFCCF value of 5-fold crossvalidation.
4. Repeat steps 1,2,3 for $k = 2$ to $Max_k = K$ so that for each *k* value, there is an optimal AFCCF value for 5-fold crossvalidation. Max_k is decided by experience and the minimum number of observations for each cluster.
5. Find the highest AFCCF value using 5-fold crossvalidation in the step 4. Then the corresponding number of clusters is the optimal number of partitions. The multicategory logit models within each cluster consist of the optimal spatial partition model.

4.3 Application to brook trout data

4.3.1 Data

The data distribution of brook trout along the eastern coast region has been introduced in the previous chapter on BCART analysis. The model we used within each cluster was logistic regression and the response variable is the binary status of brook trout population: extirpation (E) or present (P). The study area was shown in Figure 3.3. Since the status of extirpation and present was uniformly present in New Hampshire, Vermont and Maine, we excluded those observations in that study and there used only 2789 observations to build the model and analyze the data.

In the current study, on the other hand, we will use all 3337 observations since the present status has 2 categories, depressed (D) and severely depressed (SD) and is not uniform across the northeast region. Our response is categorized by the three categories of the brook trout population: extirpation (E), severely depressed (SD) and depressed (D). We use 1 for D, 2 for SD and 3 for E. Table A4.1 in the appendix of this chapter shows the distribution of the three categories within each state. This table shows that more than half of the observations in Vermont, New Hampshire and Main are depressed and the observations of SD are only half of the number of depressed. There are a total of 3337 observations. The number of observations for each category are 1083 (E), 1481 (SD) and 773(D). Among all the observations, we have 63 variables in the original dataset, as shown in Table A4.1. We will use the four variables following Zhang *et al.* (2008) in the model. The four variables are: Elevation, Road density, Agriculture and Total forest. To see how the original variables are defined, refer to Thieling (2006) and Zhang *et al.* (2008). These four variables are Box-Cox transformed first then standardized by using the mean and standard deviation of the present group.

4.3.2 Methods

In the current study, we treat the status of watershed impairment as a nominal response variable. The baseline category logit model is used to analyze the observations. The benchmark model is a baseline category logit model using all observations.

Specifically, E (extirpated, category 3) is the baseline category. D (depressed) is labeled as category 1 and SD (severely depressed) is labeled as category 2. Table 4.1 shows the parameter estimates of this benchmark model. The parameter estimates in the first row of this table is for the log odds of D over E. The second row is for the log odds of SD over E and the third is for the log odds of D over SD. The positive intercept means that the brook trout status is relatively good for the baseline logistic model. The negative elevation means that the brook trout status will degrade when elevation increases. This table shows that the estimated coefficients of elevation for both the log odds ratios are negative, with the ratio of D over E equal to -0.4986 and that of SD over E equal to -0.2645. This model implies that the higher the elevation, the worse the brook trout status. This result does not appear consistent with the expectation for brook trout. We consider this controversial result to be due to the factor that the regional differences of this spatially collected data are not considered. To deal with this problem, we apply the AFCCF criterion of 5-fold crossvalidation and run 10,000 Voronoi tessellations for each cluster size partition ranging from 2 to 7 clusters to search the optimal partition model over the entire interested space. Consequently, this optimal value accounts for the uncertainty of prediction, and is similar to the adjusted R square used in multiple regression analysis.

In the random tessellation process for the 3-category response, sometimes the observations in some cluster(s) within one tessellation can have only two or three categories with one category having a very small number of observations, say less than 15 observations. When one cluster has observations with only two categories, we use logistic regression to build the model inside by modeling the relatively 'good' category as 1. When one cluster has observations with three-category responses but with one category having less than a certain number of observations, we merge the category with the one that is close to it and then perform the logistic regression analysis using two categories. Generally, the fewer the categories, the better the prediction performance of the AFCCF criterion. Therefore, the AFCCF value for 5-fold crossvalidation for logistic regression is higher than that for multilogit regression. We found that the minimum number of observations assigned to each cluster will impact the final decision on the optimal partition model in this situation. If the minimum number of observations

assigned to each cluster is small, say less than 5% of total observations, the AFCCF criterion using 5-fold crossvalidation will pick a larger number of partitions than expected for the optimal partition. This is because the AFCCF optimal procedure will always pick partitions with high AFCCF values for individual clusters. Clusters having only two categories or clusters with three categories but with one category having a small number of observations tend to have higher AFCCFs. If the minimum number of observations assigned to each cluster is large, say 12% of total observations, the AFCCF criterion of 5-fold crossvalidation will pick a smaller number of partitions as the optimal partition. Similarly, this is because the clusters with observations having only two categories or clusters having three categories but with one category having a small number of observations are not easily obtained in this study. Currently we are interested in finding the partitions with the pattern of 3-category distribution without excluding the possible cluster(s) having observations of only two categories. After performing several simulation tests, we choose to use 9% of total observations as the minimum observation assigned to each cluster.

Table 4.1 Benchmark multilogit baseline model of brook trout data.

Parameter estimates for multilogit models based on all observations							
Category			Int	Elev	Road density	Ag	Total forest
E(3)	SD(2)	D(1)					
1083	1481	773	0.272(1/3)*	-0.499*	-1.292*	-0.673*	0.353*
			1.032(2/3)*	-0.265*	-0.443*	-0.504*	0.313*
			-0.761(1/2)*	-0.234*	-0.849*	-0.169	0.039

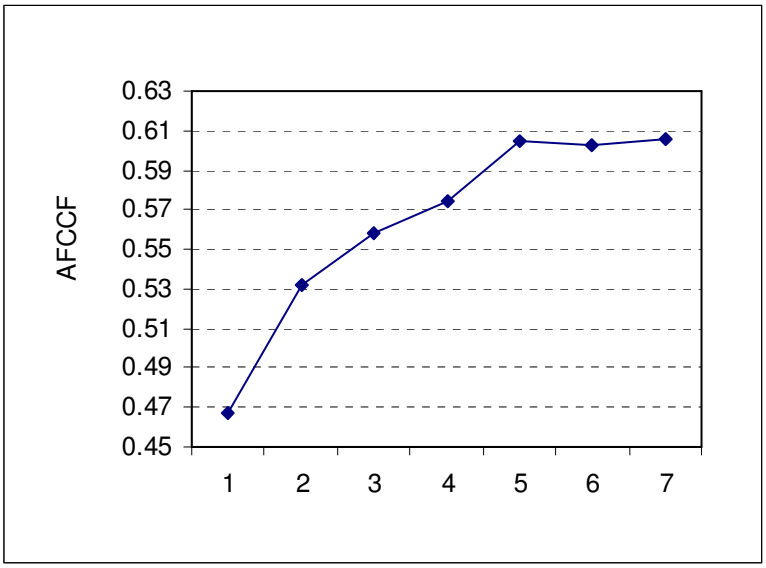
* significance level $\leq$ 0.005

4.3.3 Results for spatial partition modeling approach

The clustering result for the optimal 2- to 7-cluster partitions on the corresponding regional map is shown in Figure A4.1 in the appendix. The minimum number of observations within each cluster was set at 9% of the total observations, giving the minimum observations of 301. Figure 4.1 plots the relationship of the AFCCF values of 5-fold crossvalidation versus the number of partitions. It shows that the AFCCF value of 5-fold crossvalidation increases quickly as the number of partitions starts to increase

from size 2. It peaks for a 5-cluster partition. Then it starts to swing with a very small variation after 5-cluster partition. In this study, the AFCCF values of 5-fold crossvalidation for the partition with size larger than 5 are not always smaller than that of the 5-cluster partition. This is because we obtain more clusters, which have observations with two categories or have 3 categories but with one category having less than 15 observations. In this situation, we chose the 5-cluster partition as our optimal partition for the multicategory response models. Table A4.2 shows the optimal AFCCF values of 5-fold crossvalidation for different size partitions and the corresponding AFCCF values of 5-fold crossvalidation for each cluster within the optimal partition. In the table, the number of partitions ranges from 2 to 7. It is shown that all values of the AFCCF with crossvalidation within the cluster are greater than 0.4674, the benchmark value. This implies that the different size partition models result in considerably better prediction than the benchmark model.

Figure 4.1 AFCCF values of 5-fold crossvalidation versus number of partitions.



4.3.4 Optimal partition results: 5-cluster partition model

Figure 4.2a shows the result of the 5-cluster partition and Figure 4.2b shows the relative locations of states where the samples are present. Combining the two figures together yields the location of each cluster. The black region in the figure represents the

first cluster. It is located in the southern Appalachian area. This region has a high rate of wrongly classifying “extirpated” as “depressed” or “seriously depressed”. This area is a mountain area: it has lower agricultural activities, higher percentage of total forested lands and is very favorable to brook trout. The second cluster is given by the red region which includes all samples from western New York, Pennsylvania, and northeastern West Virginia where the Allegheny-Monongahela River Basin coal mining area is located. The third cluster, given by the green region, is located along the border of Virginia and West Virginia and extends to the southeastern corner of New York. It is within the northern Appalachian area. The fourth (blue region) and fifth (lake blue region) clusters cover Maine, New Hampshire, Vermont, Massachusetts, Eastern New York and Connecticut. Those two regions have a very high elevation and the temperature is the coolest among the five cluster regions, thus they should have the best brook trout status.

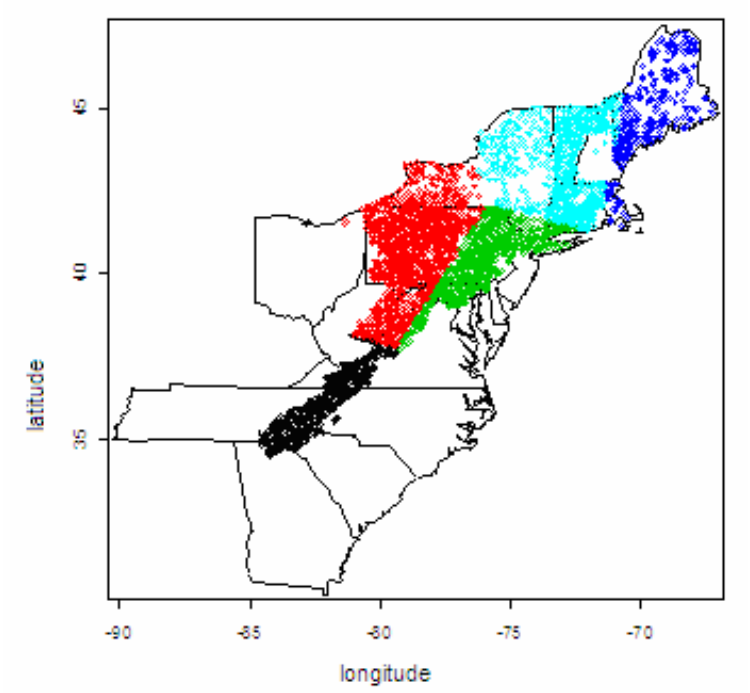
The parameter estimates for a 5-cluster multilogit partition model are summarized in Table 4.2. There are differences in both magnitudes and signs of these coefficient estimates for these 5 disjoint regions. Based on this table, we observe the following: 1) the parameter estimates for elevation in the benchmark model are all negative. When we look at the estimates obtained from a 5-cluster multilogit partition model, elevation has a uniformly positive effect on brook trout status. Elevation is partly an indicator of the water temperature within the region. The higher the elevation, the better the resulting brook trout status is. The positive elevation coefficients says that when the elevation increases, the log odds ratio of D (depressed) over E (extirpated) and SD (seriously depressed) over E will increase. In other words, the status of the brook trout will get better when elevation is higher and temperature is lower. The partition result makes a lot more sense and is consistent with the natural properties of the brook trout. 2) The intercept terms are different for each cluster. For the clusters using the multicategory logit model, the exponential of the intercept represents the probability of the mean odds of each category over the baseline category (E) at sites that have zeros for all stressor variables. For the cluster(s) using logistic regression, the exponential of the intercept represents the probability of the mean odds of the relatively ‘good’ category over the

relatively 'bad' category at sites that have zeros for all stressor variables. From Table 4.2, we see a trend for these intercept estimates. The estimates change from negative to positive and increase when the locations of regions/clusters change from south to north. Cluster 4 (the blue region) has a large positive intercept of 2.896. This region has the smallest extirpation rate (less than 15 sites) and includes only Maine and the eastern part of Massachusetts. The majority of sites in this region are not seriously depressed. Cluster 5 (the lake blue region) has the largest positive intercepts of 4.314 and 4.401. Since the difference of these two intercepts is very small and not significant (pvalue=0.658), we can say that the status of depressed and that of seriously depressed are equivalent in this region. Cluster 1 (black region) is located in the southern Appalachian region. Considering the conditions in this region, it is a good place for brook trout to live. But in our model, we have large negative estimates of intercepts (-3.012 for D over E and -2.575 for SD over E). This implies that extirpation is the dominant status of brook trout population in this region. From further research, we find that past restoration of exotic rainbow trout populations is the major reason for the brook trout population loss in this region. Although this factor was not included as a stressors, the effect was detected by the models abnormal intercept estimates. What we find for cluster 1, cluster 4 and 5 in terms of intercept estimates are consistent with that of Zhang *et al.* (2008). Cluster 2 (the red region) is a region with a coal mining effect. The odds ratio of SD over E is positively significant (1.639) and the odds ratio of D over SD is negatively significant (-2.023). The seriously depressed brook trout sites are dominant in this region. This provides the resource managers with a warning sign to take action to prevent the degrading of brook trout from SD to E. The status of brook trout in the green region (cluster 3) is similar to that of cluster 1. But the estimates of the intercepts are positive (1.610 and 2.100). According to Thieling's (2006) retrospective study, the majority of sub-watersheds predicted to be "extirpated" but in fact "present" were located in this area. 3) Although all stressors have a significant effect on the brook trout status of presence in the benchmark model, it is elevation that stands out as having a significant effect in all the regions in the 5-cluster partition model. Elevation is the only stressor for cluster 5. For the other clusters, the rest of stressors either do not have an effect or are not as significant as

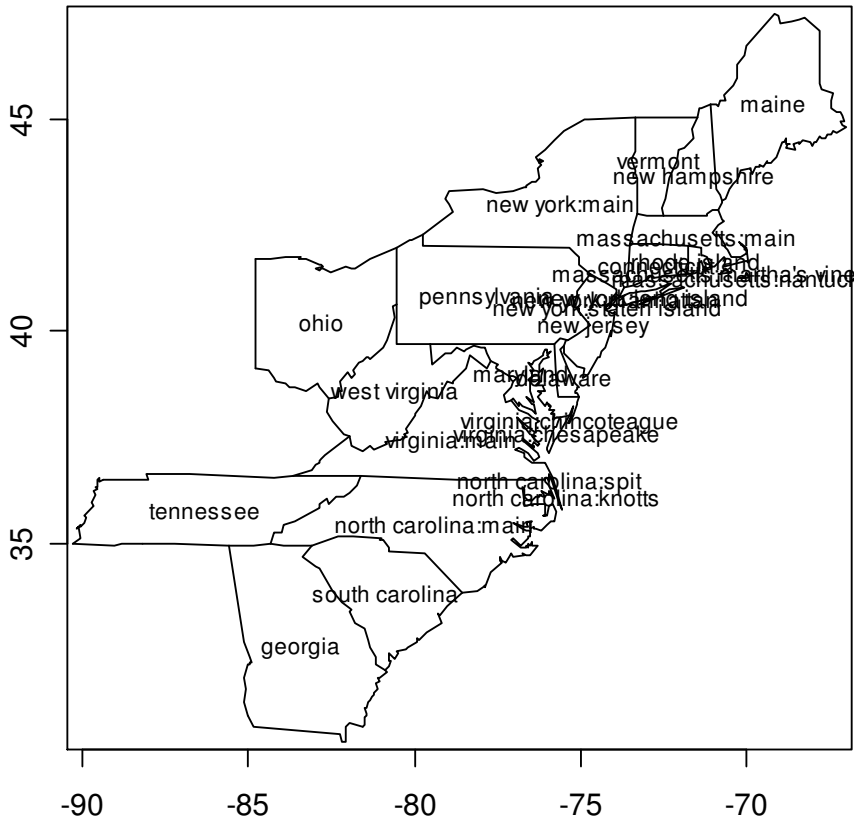
elevation. For example, in cluster 4 (the blue region), the other significant stressor is road density (with coefficient -1.168).

The previous results for cluster 1, cluster 4 and 5 are consistent with Zhang's finding with logistic regression, but with more details on the presence status for each region. This will help resource managers make more precise decisions about brook trout population maintenance and restoration. This 5-cluster partition makes more sense in terms of the three levels of brook trout status than the partition in terms of the binary brook trout status.

Figure 4.2 (a) 5-cluster partition based on the spatial partition model using AFCCF. (b) Relative locations of states where samples present.



(a)



(b)

Table 4.2 Parameter estimates for multilogit models based on 5-cluster solution.

Parameter estimates for multilogit models based on 5-cluster solution								
	Category							
Cluster	E	SD	D	Int	Elev	Road density	Ag	Total forest
1 Black	255	218	58	-3.012*	0.886*	-0.447	1.959*	1.512*
				-2.575*	1.611*	-0.300	-0.533	0.945^
2 Red	450	558	173	-0.302*	2.053*	-0.646	-1.005*	1.508*
				1.639*	1.915*	-0.823^	-0.896^	-0.163
3 Green	331	267	60	1.601*	2.921*	-0.701^	-0.599	0.811^
				2.099*	0.709*	-0.317	-0.512	0.444
4 Blue	14	105	205	2.896*	2.461*	-1.168*	0.375	0.565
5 Lake blue	33	333	277	4.314*	1.918*	-0.879	0.240	0.962
				4.401*	1.520*	-0.262	-0.043	0.376

^ significance level <0.05
* significance level <0.005

4.4 Discussion

In this chapter, we propose a multinomial category partition modeling approach by extending the binary logistic model to the multinomial logit model. The Average Fraction Correctly Classified for Fit (AFCCF) based on 5-fold crossvalidation is used as the performance measure during the search for the optimal clustering solution. Application of this method to a brook trout data set demonstrates the potential of our method for achieving better classification performance than the benchmark model. Our partition result is consistent with that of Zhang *et al.* (2008) in terms of some clusters but with more details on brook trout status.

By using this approach, we find irregular patterns in brook trout status in the Southern Appalachian area (cluster 1) that needs further investigation for other potential predictors if better prediction and management is desired. This cluster is the same as Zhang's in terms of location. We also find that the northeastern region (cluster 4 and cluster 5) has the most sites with the best conditions for brook trout populations. The majority of sites in cluster 4 are in a depressed state. The majority sites in cluster 5 are in a depressed and seriously depressed status and these two statuses have roughly the same proportions. This may provide management a warning sign that some corresponding action should be taken to prevent the SD sites from degrading into E sites although preservation and maintenance may be the correct strategy in this region. The brook trout statuses are dominated by E and SD in the cluster 2 and cluster 3 regions. It is mining activities that make the status of brook trout in the region of cluster 2 worse than that in the region of cluster 3. The coal mining activities are a major factor that degraded the brook trout status in the region of cluster 3. But this factor was not accounted in our current model. The good thing is that this region is detected by the partition approach so that further investigation of this area can be done later.

Appendix

Table A4. 1 The brook trout distribution three response categories within each state.

State	E(2)	SD(1)	D(0)	Sample size
NEW HAMPSHIRE	0	13	34	47
VERMONT	6	85	95	186
MAINE	5	88	222	315
CONNECTICUT	29	127	19	175
MASSACHUSETTS	20	80	30	130
NEW YORK	115	148	87	350
PENNSYLVANIA	444	507	134	1085
NEW JERSEY	31	24	3	58
OHIO	1	3	0	4
MARYLAND	82	42	8	132
VIRGINIA	148	56	115	319
WEST VIRGINIA	24	130	20	174
SOUTH CAROLINA	12	7	0	19
NORTH CAROLINA	95	116	3	214
TENNESSEE	18	33	3	54
GEORGIA	53	22	0	75
Total	**1083**	**1481**	**773**	**3337**

Table A4.2 AFCCF values of 5-fold crossvalidation for different size optimal partition.

Number of partitions	AFCCF of 5-fold crossvalidation within each cluster					Optimal AFCCF
	No. of Obs	E(3)	SD(2)	D(1)	AFCCF	
1	3337	1083	1481	773	0.4684	0.4674
2	Cluster1/2137	506	1012	619	0.5348	0.5324
	Cluster2/1200	577	469	154	0.5298	
3	Cluster1/1348	570	593	185	0.5789	0.5584
	Cluster2/1458	256	681	521	0.5666	
	Cluster3/531	257	207	67	0.5296	
4	Cluster1/549	262	212	75	0.5209	0.5748
	Cluster2/1037	469	442	126	0.5615	
	Cluster3/1002	56	468	478	0.5882	
	Cluster4/749	296	359	94	0.6288	
5	Cluster1/531	255	218	58	0.5332	0.6049
	Cluster2/1181	450	558	173	0.5717	
	Cluster3/658	331	267	60	0.6041	
	Cluster4/324	14	105	205	0.7600	
	Cluster5/643	33	333	277	0.5527	
6	Cluster1/612	50	349	213	0.5694	0.6031
	Cluster2/555	300	166	89	0.5687	
	Cluster3/324	5	87	232	0.7671	
	Cluster4/486	238	204	44	0.5550	
	Cluster5/474	162	236	76	0.5463	
	Cluster6/886	328	439	119	0.6123	
7	Cluster1/504	184	259	61	0.5507	0.6059
	Cluster2/306	171	124	11	0.6787	
	Cluster3/590	13	226	351	0.6083	
	Cluster4/381	132	131	118	0.4604	
	Cluster5/373	186	173	14	0.6668	
	Cluster6/301	25	163	113	0.6524	
	Cluster7/882	372	405	105	0.6239	

Table A4.3 Multicategory logit models for optimal 2- to 7-cluster partitions.

2-cluster partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
1.9828	0.4341	-1.6515	-1.2931	0.1114
2.4687	0.1672	-0.7170	-0.7968	0.2992

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 2

intercept	elev	RDKM	AG	forest
-2.7508	1.0979	-0.8177	2.0877	2.2052
-0.9997	0.9866	-0.6075	0.8581	1.0766

3-cluster Partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
-0.2411	1.9541	-0.7077	-1.005	1.1071
1.5383	1.6265	-0.7865	-0.9143	-0.2335

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 2

intercept	elev	RDKM	AG	forest
3.5414	0.9873	-2.1422	-0.9602	0.2806
3.6612	0.7929	-0.9189	-0.8503	0.2260

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 3

intercept	elev	RDKM	AG	forest
-2.8614	0.8296	-0.5329	2.0968	1.6352
-2.7698	1.7338	-0.2401	0.4393	0.8058

4-cluster Partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
-2.4803	0.7230	-0.5161	1.7291	1.3967
-2.7333	1.7252	-0.2754	0.4599	0.8372

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 2

intercept	elev	RDKM	AG	forest
0.9473	3.0539	-0.5255	-1.0175	0.6754
1.6652	0.9743	-0.2501	-0.4316	0.3042

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 3

intercept	elev	RDKM	AG	forest
5.0459	1.8322	-1.5164	-0.3610	0.6429
5.0078	1.6777	-0.3876	-0.3563	0.4410

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 4

intercept	elev	RDKM	AG	forest
-1.1582	2.3733	-0.4284	-0.1111	2.7090
1.3810	1.8883	-0.6038	-0.5815	0.5855

5-cluster Partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
-3.0117	0.8860	-0.4466	1.9586	1.5118
-2.5749	1.6119	-0.3002	0.5332	0.9453

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 2

intercept	elev	RDKM	AG	forest
-0.3020	2.053	-0.6458	-1.0053	1.5082
1.6387	1.9152	-0.8233	-0.8963	-0.1629

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 3

intercept	elev	RDKM	AG	forest
1.6099	2.9205	-0.7013	-0.5991	0.8108
2.0989	0.7087	-0.3170	-0.5122	0.4436

Coefficients estimates of LOGISTIC REGRESSION model for cluster 4

intercept	elev	RDKM	AG	forest
2.8959	2.4614	-1.1680	-0.3753	-0.5653

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 5

intercept	elev	RDKM	AG	forest
4.3143	1.9177	-0.8794	0.2401	0.9615
4.4006	1.5199	-0.2617	-0.0430	0.3757

6-cluster Partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
3.9770	1.6588	-0.5884	-0.0305	1.3705
4.1918	1.0534	-0.3711	-0.1576	0.4859

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 2

intercept	elev	RDKM	AG	forest
-0.4377	2.0269	-0.1638	-0.8866	0.2591
0.5187	0.2605	-0.5969	-0.2636	0.2195

Coefficients estimates of LOGISTIC REGRESSION for cluster 3

intercept	elev	RDKM	AG	forest
3.1374	2.6541	-1.0415	-0.4861	-0.6842

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 4

intercept	elev	RDKM	AG	forest
-4.0496	1.1869	-0.4521	2.8516	2.116
-2.6001	1.5827	-0.2256	0.4713	0.9013

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 5

intercept	elev	RDKM	AG	forest
2.0362	1.4421	-1.6621	0.0132	0.5112
2.8458	1.7876	-0.3125	-0.3311	0.3367

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 6

intercept	elev	RDKM	AG	forest
-0.4501	2.2573	-0.6100	-1.1933	1.9147
1.9621	2.5729	-0.9580	-1.0638	-0.3956

7-Cluster Partition

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 1

intercept	elev	RDKM	AG	forest
0.8496	2.2933	-0.7168	-0.0410	1.8253
2.3102	1.4518	-0.3136	-0.6518	0.2012

Coefficients estimates of LOGISTIC REGRESSION for cluster 2

intercept	elev	RDKM	AG	forest
6.8202	3.5747	-0.7441	-0.9179	0.2605

Coefficients estimates of LOGISTIC REGRESSION for cluster 3

intercept	elev	RDKM	AG	forest
0.1569	0.0255	-0.7531	0.0924	0.6707

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 4

intercept	elev	RDKM	AG	forest
-1.1305	1.7678	0.2999	-0.1625	0.8043
-0.9207	1.6788	-0.05931	-0.4018	0.5698

Coefficients estimates of LOGISTIC REGRESSION for cluster 5

intercept	elev	RDKM	AG	forest
-3.0781	1.9252	-0.3433	0.4339	0.7865

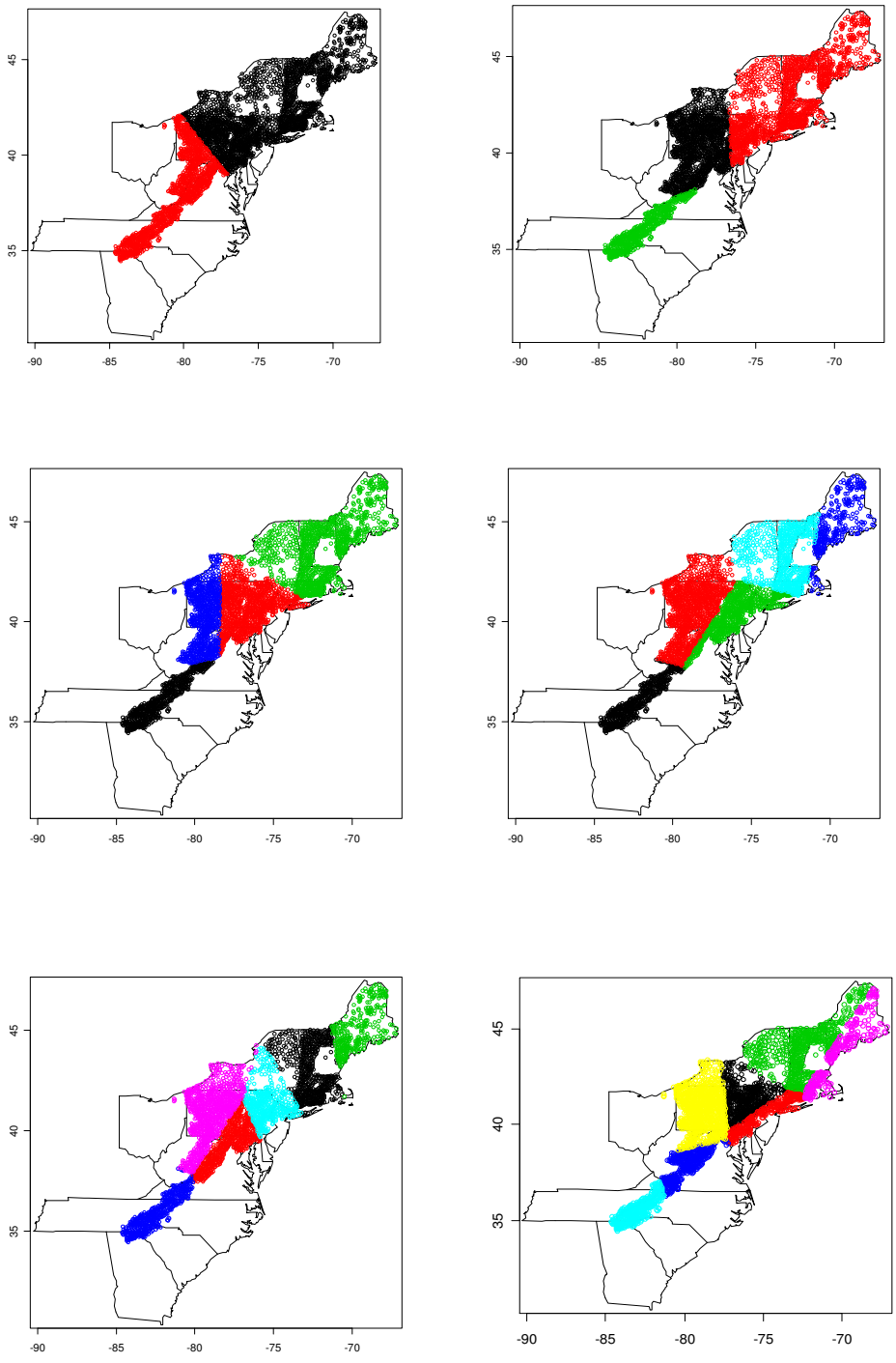
Coefficients estimates of MULTINOMIAL LOGIT model for cluster 6

intercept	elev	RDKM	AG	forest
9.0350	4.3343	-2.4606	-1.1961	-0.1312
5.6098	1.2573	-0.9060	-0.6208	0.6923

Coefficients estimates of MULTINOMIAL LOGIT model for cluster 7

intercept	elev	RDKM	AG	forest
-0.3717	2.9889	-0.7263	-1.6765	1.6266
2.1701	3.1380	-1.1025	-1.2405	-0.6729

Figure A4.1 2 to 7-cluster partitions using an AFCCF criterion.



5 Implementing the Spatial Partition Model for Multivariate Analysis

5.1 Introduction

This chapter is concerned with implementation of the Voronoi diagram based spatial partition method to a wide class of multivariate methods known as multivariate direct gradient analysis. In ecological studies, one major research interest is to find how multiple responses such as different species metrics change as the environmental stressors change. Since a multivariate relationship is harder to find than a univariate relationship, the spatial partition approach will help us to partition the large space into smaller regions and find stronger multivariate response-stressor relationships within the regions. Direct gradient methods are mostly regression based methods. Important members of multivariate direct gradient analysis methods include Canonical correspondence analysis (CCA) and Redundancy analysis (RDA). They are commonly used in the analysis of ecological data. Basically, they are multivariate regressions based on a canonical ordination technique which can reduce the multivariate dimensionality. The resulting response-stressor relationship can be displayed via biplots or triplots.

In this chapter, Redundancy analysis (RDA), Canonical correspondence analysis (CCA) and the connection between these two traditional multivariate methods will be introduced first. Then we implement the Voronoi-diagram technique to a RDA and CCA analysis and propose a weighted BIC-like model selection criterion for the Voronoi-diagram based multivariate spatial partition modeling method (MSPM). We also incorporate an AIC model selection into the partition process so that the model building process within each cluster is better than that of using the fixed stressors. We call this the Refined Multivariate Spatial Partition Modeling approach (RMSPM). The RMSPM will help us find the underlying structure of a multivariate response-stressor relationship and help resource managers to better delineate the regions they are monitoring. This method is applied to a dataset from West Virginia obtained in the years 1996-2005.

5.2 A brief review on PCA and CA

Principal component analysis (PCA) and correspondence analysis (CA) are two indirect ordination methods. They are used to summarize the variation in community composition using an ordination diagram and they are the basis for the RDA and CCA analysis. The following is the brief review on these two methods.

5.2.1 Principal component analysis (PCA)

PCA is one of the most commonly applied dimension reduction methods. It deals with a single sample of *n* observation vectors $\mathbf{y_1}, \mathbf{y_2}..., \mathbf{y_n}$ that form a swarm of points in a *p*-dimensional space. The goal in principal components analysis is to create a set of orthogonal variables (components) that are linear combinations of the original variables and maximize the variance explained in the data.

The goal is accomplished by rotating either the centered data or the centered and scaled data, so that the first principal component is the linear combination with maximal variance, the second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on. The rotation is accomplished by using either the covariance matrix or the correlation matrix of the original data. Which of these matrices is used will result in a different solution. If all the variables are not measured on the same scale it is common to use the correlation matrix.

Assuming the vector of variables $\mathbf{y} = (y_1, y_2, ...y_m)$ has been centered or centered and scaled, the axes can be rotated by multiplying observation $\mathbf{y}$ by an orthogonal matrix **A**, where $\mathbf{A'A = I}$. The *m* principal components are given by $\mathbf{z = Ay}$. **A** is chosen to maximize the variance of $\mathbf{z}$. Since $z_1, z_2..., z_m$ in $\mathbf{z = Ay}$ are uncorrelated, the sample covariance matrix of $\mathbf{z}$, $\mathbf{S_z}$ is diagonal, i.e.

$$\mathbf{S_z = ASA'} = \begin{pmatrix} s_{z_1}^2 & & 0 \\ & \ddots & \\ 0 & & s_{z_m}^2 \end{pmatrix}$$

where **S** is the sample covariance matrix of $\mathbf{y_1, y_2 ..., y_n}$. The principal components are the transformed variables $z_j = \mathbf{a}_j'\mathbf{y}, \quad j = 1, 2, ..., m$. The diagonal elements of $\mathbf{ASA'}$ are eigenvalues of **S.** The eigenvalues $\lambda_1, ..., \lambda_m$ of **S** are the variance of the principal components $z_j = \mathbf{a}_j'\mathbf{y}$ and the total variance can be completely explained by the sum of these eigenvalues, i.e., $\text{total variance} = \sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} s_{z_j}^2$. When the observations are weighted by the matrix **A**, we refer to the resulting **z** values (i.e. $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$) as scores.

5.2.2 Correspondence analysis (CA)

Correspondence analysis (CA) is a multivariate technique based on the singular value decomposition (SVD) of an *n* by *m* contingency table whose entries are counts or incidences. In the ecological study, we will consider the table as representing abundance data. Those abundances are positive integers or zero. The most common application of CA in ecology is the analysis of species data at different sampling sites. The rows and columns of the data table then correspond to sites and species, respectively. Table 5.1 is an example of an abundance data table.

Table 5.1 An example of an abundance data table

Coenagrionidae	Corbiculidae	Corydalidae	Dytiscidae	Elmidae
1	17	6	0	5
0	2	5	0	19
0	7	3	0	6
0	5	2	0	6
0	6	1	0	13
1	5	1	0	12

Correspondence analysis is similar to principal component analysis. However, it preserves, in the space of the principal axes, the Euclidean distance between profiles of

weighted conditional probabilities, which is equivalent to preserving the chi-square (χ^2) distance between the rows or columns of the contingency Table 5.1.

The following is the algebra of the method. The data is an n by m matrix of counts, **Y**. The data which may be re-expressed as proportional abundances or relative frequencies, are first transformed as $\bar{\mathbf{Q}} = \mathbf{R}^{-1/2}(\mathbf{F} - \mathbf{rc}')\mathbf{C}^{-1/2}$, where ***F*** is a relative frequency matrix and $\mathbf{F} = \mathbf{Y}/n$, $\mathbf{r} = \mathbf{Fj}$ and $\mathbf{c}' = \mathbf{j}'\mathbf{F}$. **R** and **C** are diagonal matrices containing the row and column totals, respectively. Applying the SVD results in the decomposition $\bar{\mathbf{Q}} = \mathbf{P}\Lambda\mathbf{Q}' = \sum_{r=1}^{R} \lambda_r p_r q_r$, where **P** and **Q** are orthogonal matrices, and Λ is a diagonal matrix of singular values. $\Lambda = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_k})$, where $\lambda_1, \lambda_2, ..., \lambda_k$ are non-zero eigenvalues of $\bar{\mathbf{Q}}\bar{\mathbf{Q}}'$ or $\bar{\mathbf{Q}}\bar{\mathbf{Q}}'$, *k* is the number of non-zero eigenvalues. So, we have, $\bar{\mathbf{Q}}\bar{\mathbf{Q}}' = \mathbf{P}\Lambda^2\mathbf{P}'$ and $\bar{\mathbf{Q}}'\bar{\mathbf{Q}} = \mathbf{Q}\Lambda^2\mathbf{Q}'$ or $\bar{\mathbf{Q}}\bar{\mathbf{Q}}'\mathbf{p_r} = \lambda_r^2\mathbf{p_r}$ and $\bar{\mathbf{Q}}'\bar{\mathbf{Q}}\mathbf{q_r} = \lambda_r^2\mathbf{q_r}$.

The usual chi-square statistic for testing independence of rows and columns can be expressed as $\chi^2 = n * tr[\mathbf{R}^{-1}(\mathbf{F} - \mathbf{rc}')\mathbf{C}^{-1}(\mathbf{F} - \mathbf{rc}')] = n * tr[\bar{\mathbf{Q}}\bar{\mathbf{Q}}'] = n * \sum_{i=1}^{k} \lambda_i$. So the total chi-square variance can be completely explained by the sum of those eigenvalues. The weighted average of the chi-square distance between the row profiles and their mean or between column profile and their mean is called total inertia, i.e. $Total\ inertia = \chi^2 / n = \sum_{i=1}^{k} \lambda_i$.

5.3 Relating the latent gradient to the environmental variables

An interest in community ecology is to find how a multitude of metrics of species measurements respond to external factors such as environmental variables, pollution and management etc. The two most commonly used constrained ordination techniques are Redundancy analysis (RDA) and Canonical correspondence analysis (CCA). RDA is the constrained form of Principal Component Analysis (i.e. scores are constrained to be

linear combinations of environmental variables). CCA is the constrained form of Correspondence Analysis. CCA is appropriate under a linear model, as long as one is interested in species composition rather than absolute abundances (Legendre, 1998). Since most of the discussions concering CCA also relates to RDA, the relationship between them will also be discussed.

5.3.1 Redundancy analysis (RDA)

Redundancy analysis (RDA) is the direct extension of multiple regressions to the modeling of multivariate response data. Suppose there are two tables, ***Y*** is the table of response variables with size *n* by *m,* **X** is the table of explanatory variables with size *n* by *p* ($p \le n$). In redundancy analysis, ***X*** and ***Y*** each are first centered. The canonical ordination axis corresponds to a direction, which is maximally related to a linear combination of the environmental variables **X**. A canonical axis is similar to a principal component. Two ordinations of the objects are obtained along the canonical axes: (1) linear combinations of the species variables as in PCA, and (2) linear combinations of fitted species variables which are also linear combinations of X variables. RDA preserves the Euclidean distance among objects in the matrix $\hat{\mathbf{Y}}$. Variables in $\hat{\mathbf{Y}}$ are linear combinations of the X variables.

The mathematics of RDA is typically based on the partitioned covariance matrix from the **Y** and **X** data sets.

$$\mathbf{S}_{\mathbf{Y+X}} = \left[\begin{array}{ccc||ccc} s_{y_1,y_1} & \cdots & s_{y_1,y_m} & s_{y_1,x_1} & \cdots & s_{y_1,x_p} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ s_{y_m,y_1} & \cdots & s_{y_m,y_m} & s_{y_m x_1} & \cdots & s_{y_m,x_p} \\ \hline\hline s_{x_1,y_1} & \cdots & s_{x_1,y_m} & s_{x_1,x_1} & \cdots & s_{x_1,x_p} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ s_{x_p,y_1} & \cdots & s_{x_p,y_m} & s_{x_p,x_1} & \cdots & s_{x_p,x_p} \end{array}\right] = \begin{pmatrix} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S_{XY}} & \mathbf{S_{XX}} \end{pmatrix}$$

where $\mathbf{S_{YY}}$ (*m by m*) and $\mathbf{S_{XX}}$ (*p by p*) concern the variance of the two sets of descriptors, $\mathbf{S_{XY}}$ (*p by m)* and its transpose account for covariance among the descriptors of the two groups.

Redundancy analysis is a two-step process: (1) regress each variable in **Y** on all variables in **X** and compute the fitted values, and (2) do a principal component analysis of the matrix of fitted values to obtain the eigenvalues and eigenvectors (Legendre, 1998). In multiple regression, the fitted values are computed as $\mathbf{\hat{Y} = XB}$, where **B** is the matrix of regression coefficients of all response variables **Y** on the regressors **X** and $\mathbf{B = [X'X]^{-1}X'Y}$, hence $\mathbf{\hat{Y} = X[X'X]^{-1}X'Y}$. The covariance matrix corresponding to the fitted values $\mathbf{\hat{Y}}$ is computed as $\mathbf{S_{\hat{Y}\hat{Y}}} = [1/(n-1)]\mathbf{Y'X[X'X]^{-1}X'X[X'X]^{-1}X'Y} = \mathbf{S_{YX}S_{XX}^{-1}S_{YX}}'$. To reduce the dimensionality of the solution, the fitted values, $\mathbf{\hat{Y}}$, are subjected to principal component analysis. The components are obtained from an eigen analysis i.e.,

$$(\mathbf{S_{\hat{Y}\hat{Y}}} - \lambda^* I)u = (\mathbf{S_{YX}S_{XX}^{-1}S_{YX}'} - \lambda^* I)u = 0$$

where λ^*s are the eigenvalues of the covariance matrix corresponding to the fitted values $\mathbf{\hat{Y}}$, called canonical eigenvalues. The sum of these canonical eigenvalues is the variance explained by the environmental variables. The matrix of eigenvectors **U,** containing the normalized canonical eigenvectors of size *m by m,* but contains only *min[p,m,n-1]* eigenvectors with non-zero eigenvalues. The canonical coefficients in **U** give the contributions of the variables of $\mathbf{\hat{Y}}$ to the canonical axes. The ordination of objects in the space of **X** is obtained as $\mathbf{Z = \hat{Y}U = XBU}$. It is referred to as the "fitted site scores". This ordination vectors have variances equal to the corresponding eigenvalues. If we let $\mathbf{C = BU}$, then the coefficients in the columns of matrix **C** are identical to the regression coefficients associated with the ordination scores **Z**.

5.3.2 Canonical correspondence analysis (CCA) and its relationship with RDA

Canonical correspondence analysis is similar to RDA. The difference is that it preserves the χ^2 distance as in correspondence analysis, instead of the Euclidean

distance among objects. Calculations are a bit more complex since the matrix $\hat{\mathbf{Y}}$ contains the fitted values obtained by the weighted linear regression of matrix $\bar{\mathbf{Q}}$ of correspondence analysis (CA), the transformed observation **Y,** on the explanatory variables **X.** The algorithm of CCA is to first regress the transformed **Y** onto appropriately standardized **X**'s and then compute the site and column scores by applying the Singular Value Decomposition (SVD) to the matrix of the fitted values. Thus, the obtained site scores are linear combinations of the environmental variables by construction. This technique allows one to display the site scores, species scores and environmental variables on the same ordination diagram.

In a CCA, the dependent table is not the matrix **Y** centered by variables as in RDA, rather CCA uses the matrix $\bar{\mathbf{Q}}$, the contributions to Chi-square used in CA which is $\bar{\mathbf{Q}} = \mathbf{R}^{-1/2}(\mathbf{F} - \mathbf{rc}')\mathbf{C}^{-1/2}$, where **F** is a relative frequency matrix ($\mathbf{F} = \mathbf{Y}/n$), and $\mathbf{r} = \mathbf{Fj}$ and $\mathbf{c}' = \mathbf{j}'\mathbf{F}$. **Y** is the n by m contingency table, **R** and **C** are the diagonal matrices with the row and column totals, respectively. The matrix **X** is standardized using a weight matrix, $\mathbf{D}(r_{i+})$, which is the diagonal matrix of row relative frequencies of the matrix **Y.** The mathematics of CCA is essentially the same as that of redundancy analysis, although weighted regression is used instead of a conventional multiple regression. The weights, given by $\mathbf{D}(r_{i+})^{1/2}$, are applied to matrix **X** everywhere it occurs in the multiple regression equations. So we have $\mathbf{B} = [\mathbf{X}'\mathbf{D}(\mathbf{r}_{i+})\mathbf{X}]^{-1}\mathbf{X}'\mathbf{D}(\mathbf{r}_{i+})^{1/2}\bar{\mathbf{Q}}$ and $\hat{\mathbf{Y}} = \mathbf{D}(\mathbf{r}_{i+})^{1/2}\mathbf{XB} = \mathbf{D}(\mathbf{r}_{i+})^{1/2}\mathbf{X}[\mathbf{X}'\mathbf{D}(\mathbf{r}_{i+})\mathbf{X}]^{-1}\mathbf{X}'\mathbf{D}(\mathbf{r}_{i+})^{1/2}\bar{\mathbf{Q}}$. Thus $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{S}_{\mathbf{Q}\bar{\mathbf{X}}}\mathbf{S}^{-1}_{\mathbf{XX}}\mathbf{S}'_{\mathbf{Q}\bar{\mathbf{X}}}$.

To obtain eigenvalues and eigenvectors, CCA is completed using the eigenvalue equation

$$(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}} - \lambda^{**}\mathbf{I})\mathbf{u} = (\mathbf{S}_{\bar{\mathbf{Q}}\mathbf{X}}\mathbf{S}^{-1}_{\mathbf{XX}}\mathbf{S}'_{\bar{\mathbf{Q}}\mathbf{X}} - \lambda^{**}\mathbf{I})\mathbf{u} = 0$$

where λ^{**} is the eigenvalue corresponding to the covariance matrix of fitted values $\hat{\mathbf{Y}}$, called the canonical eigenvalue. The sum of these canonical eienvalues is the variance explained by the environmental variables and is called constrained inertia.

The maximum number of non-zero eigenvalues and corresponding eigenvectors that may be obtained from a RDA and a CCA with a matrix of response variable **Y** $(n \times m)$ and explanatory variables **X** $(n \times p)$ are shown in Table 5.2 (Legendre, 1998).

5.3.3 Data analysis and ordination diagram

Prior to a RDA, the data matrices will be transformed. The response variables **Y** will be centered on their means, or standardized by column if the variables are not dimensionally homogeneous. The explanatory variables **X** will be centered on their respective means. They may also be standardized to remove the scale effect of the physical dimension of the explanatory variables and turn the regression coefficients into standard regression coefficients, but it is not necessary for a valid RDA. For a CCA, the response variables **Y** will be transformed to $\bar{\mathbf{Q}}$ and the weights, given by $\mathbf{D}(r_{i+})^{1/2}$, are applied to the matrix **X** everywhere it occurs in the multiple regression equations. The SVD decomposition is then carried out on the matrix $\hat{\mathbf{Y}}$ to produce $\hat{\mathbf{Y}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$, where the columns of **U** are corresponding eigenvectors of $\hat{\mathbf{Y}}\hat{\mathbf{Y}}'$, columns of **V** are corresponding eigenvectors of $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix containing square roots of the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ of $\hat{\mathbf{Y}}\hat{\mathbf{Y}}'$ (or $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$). The package 'vegan' in the R program is based on CANOCO. The standardization is automatically performed for the explanatory variables when computing a RDA and a CCA.

Table 5.2 The maximum number of the non-zero eigenvalues and corresponding eigenvectors (from Legendre 1998).

	Canonical eigenvalues and eigenvectors	**Non-canonical eigenvalues and eigenvectors**
RDA	Min[*p*,*m*,*n*−1]	Min[*p*,*n*−1]
CCA	Min [(*p*−1),*m*,*n*−1]	Min[(*p*−1),*n*−1]

An ordination diagram is used to help visualize the main structure of multivariate relations in two or three dimensions. In an RDA/CCA analysis, the complete canonical ordination diagram consists of three sets of points: sites, species and environmental variables. We call the graph a triplot if all three sets of points are plotted on the same graph.

The fitted site coordinates (in the space of $\mathbf{X}$'s) are $\mathbf{U} = \hat{\mathbf{Y}}\mathbf{V}\mathbf{\Lambda}^{-1}$. Column scores or species coordinates represent the columns of $\hat{\mathbf{Y}}$. The response/species coordinates are obtained from the first few columns of the right-hand singular vectors in $\mathbf{V}$. $\mathbf{U}$ can be paired with $\mathbf{V}$ to form a biplot. The coordinates for the environmental variables in $\mathbf{X}$ can be added on the same diagram to form a triplot. These coordinates are calculated as correlations of the fitted site coordinates and the columns of environmental variables in $\mathbf{X}$.

Using the singular value decomposition of $\hat{\mathbf{Y}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$, the two dimensional representation of $\hat{\mathbf{Y}}$ is

$$\hat{\mathbf{Y}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' = \mathbf{Z}\mathbf{A}' \cong \mathbf{Z}_2\mathbf{A}_2' = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \\ z_{n1} & z_{n2} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \cdots & a_{m1} \\ a_{12} & a_{22} \cdots & a_{m2} \end{pmatrix}$$

The Euclidean distance between the two rows of $\mathbf{Z}_2$ is approximately equal to the distance between two points from the data matrix $\hat{\mathbf{Y}}$. $\mathbf{Z}_2$ is the site scores from the first two axes. The angle between the two columns of $\mathbf{A}_2$ is approximately the correlation of the two response variables. $\mathbf{A}_2$ is the species/response score of the first two axes. Alternative factorings of $\hat{\mathbf{Y}}$ provide different scaling for a triplot and may be useful for plotting different characteristics.

In this plot, sites are indicated by points while both species and environmental variables are indicated by arrows. The imaginary axis defines the direction along which

the values of the estimated responses/species metrics change. The arrows point in the direction of the maximum variation in the species abundance/responses and environmental variables. Arrows pointing in roughly the same direction indicate high positive correlation and arrows pointing in opposite directions indicate high negative correlation. Species and environmental variables with long arrows are most important in the analysis. The longer the arrow, the greater the correlation. Also, the length of arrows is proportional to the correlation between ordination and environmental variables. The sites should be perpendicularly projected onto the arrows to predict the value of the variable.

5.3.4 An example

The following is an example of a CCA with a triplot using the famous hunting spider data from ter Braak (1986). This example shows how canonical correspondence analysis can be used to detect species-environmental relationships. The data set concerns the distribution of spiders in a dune area. The original data consists of the counted abundances of twelve species captured at 100 sites in a dune area in the Netherlands. Six environmental variables were measured at 28 sites of the 100 sites. These were water content, bare sand, moss cover, light reflection, fallen twigs and herb cover. We use the observed abundances at the 28 sites where the environmental variables were monitored. Species data is transformed by taking square roots. The 12 species are given in Table 5.3.

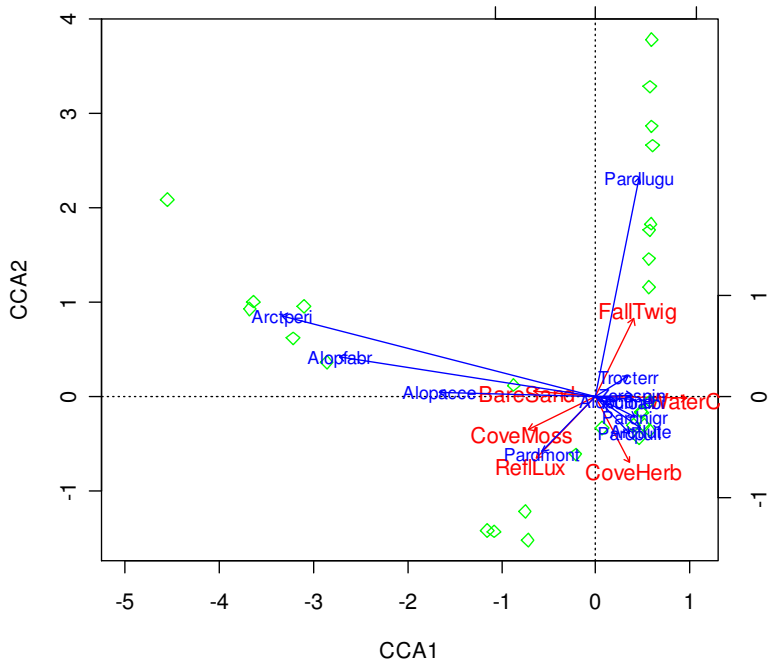
Table 5.3 Twelve spider species in the Netherlands example.

No.	Species	No.	species
1	Arctosa lutetiana	7	Trochosa terricola
2	Pardosa lugubris	8	Alopecosa cuneata
3	Zora spinimana	9	Pardosa monticola
4	Pardosa nigriceps	10	Alopecosa accentuata
5	Pardosa pullata	11	Alopecosa fabrilis
6	Aulonia albimana	12	Arctosa perita

The triplot in Figure 5.1 shows a correlation between species, environmental variables and samples. Red lines pointing in the same direction indicate that the corresponding explanatory variables are correlated with each other. Examples are moss

cover and light reflection. Long lines are more important than the short ones. Lines pointing in opposite directions are negatively correlated, see for example the lines for water content and bare sand. It doesn't come as a surprise that these two are negatively correlated. Lines with an angle of 90 degrees indicate that the two variables are uncorrelated, for example, moss and herb cover. The same interpretation holds for the species. For example, *A. perita* (denoted by Arctperi) and A. fabrilis (Alopfabr) are highly correlated. One can see that the first CCA axis can be associated with the aridness of the area. The second axis is related to the type of the soil cover in the area of habitat.

Figure 5.1 Triplot for Hunting Spider data.



5.4 Implementing spatial partition model for CCA and RDA

5.4.1 Multivariate spatial partition modeling (MSPM)

The Voronoi diagram partition method in chapter 2 proposed the method of clustering sites to maximize the strength of cluster-wise stressor-response relationships within clusters. Here in multivariate analysis, we are concerned with the strength of cluster-wise multivariate stressor-response association. The Voronoi diagram is used to

randomly assign sites to one of K clusters based on spatial measurements (latitude and longitude or width and elevation). In this chapter, one major research interest is to find how the responses such as different species' metrics change as the environmental stressors change. We propose a weighted BIC-like criterion to find the underlying multivariate structure association over entire region. The weighted BIC-like criterion will be developed in the next two sections in detail. The extension of the Voronoi diagram partition method to multivariate canonical analysis such as CCA and RDA is called multivariate spatial partition modeling (MSPM). The detailed steps in this method are as follows,

1. Choose the model selection criterion for the analysis of interest.
2. Partition the two dimensional spatial region of the data into k non-overlapping clusters using the Voronoi diagram technique.
3. Perform an RDA/CCA within each cluster using the fixed number of stressors and calculate the value of the criterion for this k-clusters model.
4. Repeat steps 2 and 3 enough times so that the optimal k clusters are found.
5. Repeat steps 2, 3, 4 for $k = 2$ to $Max_k = K$ so that for each k value, there is an optimal value. Max_k is decided by experience and minimum number of observations for each cluster .
6. Decide the optimal clustering (model) according to the result from step 5.

The MSPM approach can be adjusted to detect the hotspot region over a large study region. The hotspot detection approach will be introduced and developed in the next chapter.

5.4.2 Log likelihood for RDA/CCA from reduced rank regression

Let $\mathbf{Y}$ be the n *by* m matrix of response variables and $\mathbf{X}$ be the n *by* p matrix of stressors. Each observation is assumed to have been taken from a single site. Let h_1, h_2 be the two dimensional spatial measurement variables associated with each observation, the clusters of all sites are defined by them such that observations have a similar dominant

predictor-response relationship within a cluster and have different relationship between clusters. Let n_k be the number of observations placed in the k^{th} cell, where $n_1 + n_2 + ... + n_K = n$.

CCA and a RDA are special cases of reduced rank regression (RRR). The model for reduced rank regression is $\mathbf{Y = XM + E}$, where $\mathbf{M}$ is a *p* by *m* matrix of coefficients. It has rank $r < \min(m, p)$. $\mathbf{E}$ is the error term and is assumed to be uncorrelated across rows and normally distributed as $\mathbf{E \sim N(0, \Sigma)}$. Davies and Tso (1982) showed that the weighted error sum of squares allows the following orthogonal decomposition: $\left\|\mathbf{(Y - XM)\Gamma}\right\| = \left\|\mathbf{(Y - \hat{Y})\Gamma}\right\| + \left\|\mathbf{(\hat{Y} - \hat{Y}_r)\Gamma}\right\|$, where $\hat{\mathbf{Y}}$ is the matrix of the fitted values from ordinary least squares (OLS) regression and its SVD is $\hat{\mathbf{Y}} = \mathbf{U\Lambda V'}$; $\hat{\mathbf{Y}}_\mathbf{r}$ is the RRR fit, and $\mathbf{\Gamma}$ is a symmetric positive definite matrix. By using the Eckart-Young theorem, the minimum of this quantity is achieved by retaining the first *r* columns of the orthogonal matrices **U** and **V** with their associated singular values λ_i, so $\hat{\mathbf{Y}}_\mathbf{r} = \mathbf{U}_\mathbf{r}\mathbf{\Lambda}\mathbf{V}_\mathbf{r}'$.

As noted in ter Braak (1987), the solution to the RRR problem is equivalent to the RDA solution when $\mathbf{\Gamma}$ is the identity matrix which corresponds to $\mathbf{\Sigma = I}\sigma^2$ and equivalent to the CCA solution when $\mathbf{\Gamma} = \hat{\mathbf{\Sigma}}_{ols}^{-1/2}$ which corresponds to $\mathbf{\Sigma}$ being unspecified. The likelihood for the reduced rank regression with normal errors $\mathbf{\Sigma}$ is:

$$\ln L = \text{constant} + n/2 * \ln\left|\mathbf{\Sigma^{-1}}\right| - (1/2\hat{\sigma}^2)\mathbf{tr\{(Y - XM)\Sigma^{-1}(Y - XM)'\}}$$

Lipkovich (2003) summarized the likelihood derivation when $\mathbf{\Gamma}$ takes several different forms. The following is the brief summary on 3 related forms.

(1) $\mathbf{\Gamma = I}, \mathbf{\Sigma = I}\sigma_0^2$ where σ_0^2 is known

(2) $\mathbf{\Gamma = I}, \mathbf{\Sigma = I}\sigma^2$ where σ^2 is unknown and need to be estimated from data and

(3) $\mathbf{\Gamma = \hat{\Sigma}_{ols}^{-1/2}}$

In situation (1), the maximized twice log likelihood is $2\ln L = \text{constant} + \sum_{i=1}^{r} \lambda_i^2$, where λs are the singular values of the matrix of the fitted values $\hat{\mathbf{Y}}_{ols}$. When $\mathbf{M}$ is of full rank, this reduces to the trace of the corresponding matrix $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$. In practice, situation (2) is of interest. Since σ^2 can be estimated by $\hat{\sigma}^2 = (nm)^{-1} tr\{(\mathbf{Y} - \mathbf{X}\hat{\mathbf{M}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{M}})\}$, the maximized twice log likelihood is $2\ln L = \text{constant} + n\ln\hat{\sigma}^2 = \text{constant} + n\ln\left\{ n^{-1}\text{trace}(\hat{\mathbf{\Sigma}}_{\mathbf{ols}}) + \sum_{i=r+1}^{m} \lambda_i^2 \right\}$, where $\hat{\mathbf{\Sigma}}_{\mathbf{ols}}$ is the ordinary least squares estimate of $\mathbf{\Sigma}$ using $\mathbf{Y} - \hat{\mathbf{Y}}$. The λ s are ordered singular values of $\hat{\mathbf{Y}}_{ols}$. $\sum_{i=r+1}^{m} \lambda_i^2$ is the sum of the last $m - r$ singular values. For situation (3), when the covariance matrix $\mathbf{\Sigma}$ is unknown, the maximized twice log likelihood is $2\ln L = \text{constant} + n\ln\left|\mathbf{\Sigma}^{-1}\right| = \text{constant} + n\sum_{i=1}^{r} \ln(1+\lambda_i)^{-1}$. The MLE estimate for $\mathbf{\Sigma}$ was given by Davies and Tso (1982) using Wilk's lambda representation of the likelihood ratio statistic $\Lambda = \prod_{i=1}^{p} (1+\lambda_i)^{-1} = \frac{|E|}{|E+H|}$, the λs now are the eigenvalues of the matrix $\mathbf{E}^{-1}\mathbf{H} = [(\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{ols}})'(\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{ols}})]^{-1}\mathbf{Y}'\mathbf{Y}$.

5.4.3 Weighted BIC criterion for one tessellation in RDA analysis

For a RDA analysis, the BIC-like criterion can be expressed as $BIC = \ln L - penalty$ and we want to maximize this quantity to choose our partition model. For the k^{th} cluster in the multivariate RDA analysis, this quantity is approximated by $BIC_k = n_k \ln\left\{ \mathrm{n_k}^{-1} tr(\hat{\mathbf{\Sigma}}_k) + \sum_{i=r+1}^{m} \lambda_{ik}^2 \right\} - p \ \ln n_k$

For one tessellation with K clusters, we define a weighted BIC as $BIC_W = \sum_{k=1}^{K} n_k * BIC_k / N = \sum_{k=1}^{K} n_k \left[n_k \ln\left\{ \mathrm{n_k}^{-1} tr(\hat{\mathbf{\Sigma}}_k) + \sum_{i=r+1}^{m} \lambda_{ik}^2 \right\} - p \ln n_k \right] / N$. When we use a

full rank model, $r = m$, and the $\sum_{i=r+1}^{m} \lambda_{ik}^2$ part disappears. Usually in our data analysis, the minimum of the observations within a chosen cluster is greater than $\max(p, m)$. When the number of response variables m is less than that of the environmental variables, p, we have the full rank situation, i.e. $r = m$, and the weighted BIC criterion is

$$BIC_W = \sum_{k=1}^{K} n_k \left[n_k \ln\left(n_k^{-1} tr(\hat{\Sigma}_k) \right) - p \ln n_k \right] / N . \quad (1)$$

When the number of response variables, m, is greater than that of the environmental variables (stressors), p, we have the weighted BIC criterion

$$BIC_W = \sum_{k=1}^{K} n_k \left[n_k \ln\left\{ n_k^{-1} tr(\hat{\Sigma}_k) + \sum_{i=r+1}^{m} \lambda_{ik}^2 \right\} - p \ln n_k \right] / N . \quad (2)$$

5.4.4 Decide the number of clusters for the optimal partition

The weighted BIC value usually increases dramatically as the partitioning procedure starts then will begin to level off as the number of partitions increases. When the number of partitions is close to a certain size, for example *k*, the weighted BIC value starts to increase very slowly from the *k+1*th cluster partition onward. We call the *k*-cluster partition the partition of change point for the weighted BIC value. The minimum number of observations assigned to each cluster, $Min_k obs$, has a certain effect on the relationship of the weighted BIC value vs. the number of clusters. The smaller the value of $Min_k obs$ relative to the total number of observations, the larger the number of partitions or clusters, k, . But within an adequate range of $Min_k obs$, the partition change point for the weighted BIC value will not change by much even if the $Min_k obs$ is small. If the $Min_k obs$ is large relative to the total number of observations, say 15% of the total number of observations, then the number of partitions k will be relative small. In this case, we may not be able to observe the partition change point for the weighted BIC value by plotting the relationship of the weighted BIC values versus the number of clusters. Whenever if we choose an adequate value of $Min_k obs$, the weighted BIC value will increase dramatically before the number of partitions reaches a certain number. Then

it will increase slowly after that. By adequate, we mean that $Min_k obs$ should be large enough to avoid the model overfitting within each cluster. The change of the weighted BIC value can be described by looking at $Diff_{WBIC_k} = BIC_{W(k+1)} - BIC_{W(k)}$, the difference of the weighted BIC values between the *k*-cluster and *k+1*-cluster partitions. We can plot the relationship between $Diff_{WBIC}$ and the number of partitions to see the change in the weighted BIC value across different number of partitions to find the partition of change point, or the optimal partition, just like how a scree plot in principal component analysis is used to decide the number of components to retain. We can also use a 95% cut point to help us to make the decision. Let $Ratio_k = Diff_{WBIC_k} / BIC_{benchmark}$, then $Ratio_k$ is the change of weighted BIC value over the benchmark model BIC value for the *k*-cluster partition. If $Ratio_k$ for the *k*-cluster partition is less than 5%, then we will use the corresponding number of partitions as our optimal number of partitions.

5.4.5 Refined multivariate partition model (RMSPM)

In order to get a better model within each cluster for one partition/tesselation, it is very natural to incorporate a variable selection process within the partition model selection procedure. We will use the analog of the AIC variable selection method suggested by Godinez-Dominguez and Freire (2003) to do variable selection at the same time as we partition the data so that the partition with the highest weighted BIC is selected. In other words, we want $tr(\hat{\Sigma}_k)$ in formula (1) or (2) to be a maximum for each cluster within a specific partition so that the weighted BIC value will be as large as possible. The steps in this RMSPM method are similar to that of MSPM. Its procedure is as the follows,

1. Partition the two dimensional spatial region of data into *k* non-overlapping clusters using the Voronoi diagram technique.
2. Perform a RDA/CCA within each cluster with variable selection and calculate the value of the BIC-like criterion for this *k*-cluster model.
3. Repeat steps 2 and 3 a sufficient number of times such that the *k*-cluster partition with the largest weighted BIC-like value is found.

4. Repeat steps 2, 3, 4 for $k = 2$ to $Max_k = K$ so that for each k value, there is an optimal value. The Max_k is decided by experience and a minimum number of observations for each cluster.
5. Decide the optimal partition (model) according to step 4 by using a scree plot of $Diff_{WBIC}$ vs. the number of clusters and/or a 5% cut point.

5.5 Application to WV data using refined multivariate partition model (RMPM)

5.5.1 The West Virginia dataset

The West Virginia data set includes 4216 samples of benthic macroinvertebrates sampled from 1996 to 2005. There are six multi-metric indexes which are overall indicators of the stream condition. They are the Total Taxa (Total_Taxa), the count of Ephemeroptera, Plecoptera and Trichoptera Taxa (EPT_Taxa), the Percentage of EPT (P_EPT), the percentage of Chironomidea (P_Chiro), the percentage of the Two Dominant Taxa (P_2Dom) and a family level Hilsenhoff Biotic Index (HBI). The seventh metric, WVSCI, is the combination of the previous 6 metrics. It will be used to classify each observation into a reference site (site of good condition where WVSCI>78) or a test site (site of bad condition where WVSCI<78) (Green, 2000) in our analysis and will not be included in the partition procedure.

According to Green (2000), the metric Total Taxa measures the number of families in the sample. It generally decreases with increasing stream degradation. The metric EPT_Taxa measures taxa richness in three insect orders known to be generally sensitive to changes in water quality (Ephemeroptera, Plecoptera and Trichoptera). It generally decreases with degrading stream conditions. The metric P_EPT is based on the proportion of individuals in the sample that belong to the EPT orders. We generally expect that in healthy streams, a high percentage of the total organisms present should belong to the EPT orders. The metric P_Chiro is based on the proportion of individuals in the sample that belong to the family Chironomidae. This metric generally increases with

degrading stream condition. The Hilsenhoff Biotic Index (HBI) metric can be thought of as an average organic pollution tolerance value for the sample, weighted by the abundance of organisms. This metric increase with degrading stream conditions and is often used as a general indication of stress. The metric P_2Dom is based on the proportion of individuals in the sample that belong to the two most dominant taxa. In healthy streams, there are generally several families, with the individuals evenly distributed among the different families. As stream degradation occurs, more individuals are concentrated in fewer, more tolerant families, and this metric generally increases. In summary, metrics Total_Taxa, EPT_Taxa, P_EPT are positive in the sense that high values indicate good conditions while the metrics P_Chiro, P_2Dom and HBI are negative with low values representing good conditions. The six metrics were aggregated into an index WVSCI which has a 100 point scale (Green, 2000)

There are eleven rapid habitat assessment variables: Cover, Embed, Velocity, Alter, Sediment, Riffl_Sinu, Chanflow, BnkStbTot, BnkVegTot, RipVegTot and total which is the sum of previous ten parameters. Those ten variables were scored from 1 to 20 and the higher the score, the better the habitat condition. Four field Chemical/Physical measurements: conductivity (Conduct), PH, temperature (Temp) and dissolved oxygen (DO) were measured. Also we have fifteen laboratory Chemical/physical measurements. They are Fecal, Acid_Hot, Alkalinity, Hardness, Sulfate, Chloride, Tot_Phos, NO2_NO3_N, Al_Tot, Ca_Tot, Cu_Tot, Fe_Tot, Mg_Tot, Mn_Tot, Zn_Tot. For the meaning of these habitat and chemical variables see Table A5.1 in the appendix for the description. The location variables for each observation are latitude and longitude. Figure A5.1 in the appendix gives the yearly sample distribution for the West Virginia data from 1996 to 2005.

5.5.2 Data manipulation before analysis

There are two richness metrics, Total Taxa, EPT_Taxa, and three percentage metrics, P_EPT, P_Chiro and P_2Dom, among the six multi-metric indexes. Some metrics are redundant with others. Total Taxa is highly correlated with metric EPT Taxa (with correlation 0.8252), metric P_EPT is highly correlated with metric P_Chiro (with

correlation -0.8278). To reduce the effect of extreme values, we used arcsin($\sqrt{y/100}$) for percent metrics. The richness metrics were not transformed. Since a randomized selection procedure is used to select the site, the environmental variables exhibited skewed distribution (Griffith *et al*., 2001). The chemistry and habitat measurements were Box-Cox transformed to reduce the skewed distribution for some variables after adjustment for missing values. The RDA/CCA package in R or CANOCO was used to standardize the environmental variables; this removes the effect of differences in measurement units.

From Table A5.1 in the appendix we find that the missing values are mainly from the lab chemical variables. If we delete all the observations with missing values in any of the variables, then only 1657 observations remain. This will hurt our final analysis since only approximately 1/3 of the observations will be used. To deal with this, we will group the lab chemical variables into several index variables which combine variables that are similar in degree of effect on macroinvertebrates in the watershed and average these similar effect variables. This results in half of the original observations.

Sulfate and chloride are both anions and are measured as the dissolved substances, so we combined these two variables into a new variable labeled anion. Aluminum, copper, iron, magnesium, manganese and zinc are all heavy metals and we combined these into a new variable labeled metal. Total phosphorus and nitrates are both nutrients and we combined them into a nutrient variable labeled nutri. Alkalinity and hardness were combined into ak_hard variable. Using the derived metal variable as the example, the combination was computed to obtain $metal_j = \frac{1}{p_i}\sum_{i=1}^{p_i} z_{ij}$

where $z_{ij} = \frac{X_{ij} - \bar{X}_{i\_ref}}{STD_{i\_ref}}$. X_i stands for one variable among aluminum, copper, iron, magnesium, manganese and zinc. X_{ij} is the non-missing values for that variable. $\bar{X}_{i\_ref}$ is the mean of the reference observations for variable X_i and STD_{i_ref} is the standard deviation of the reference observations for variable X_i. So z_{ij} is the standardized

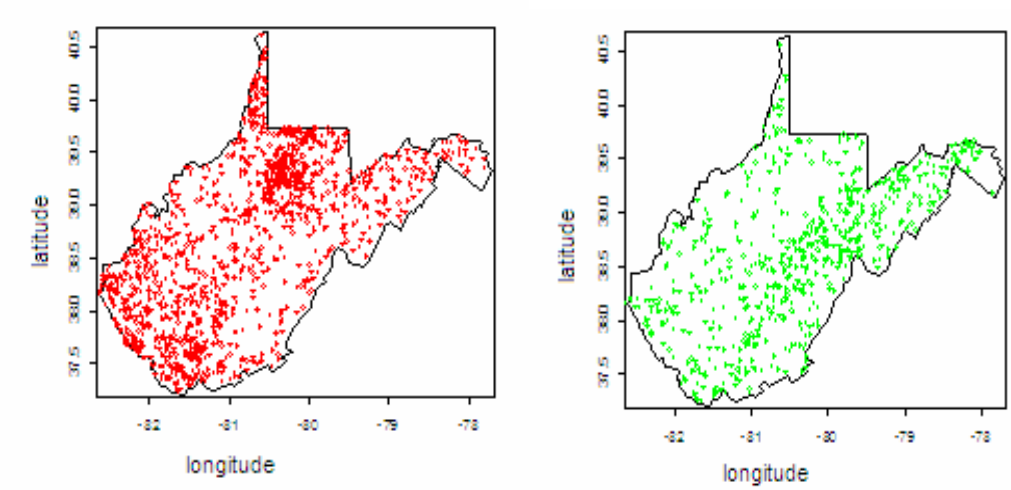
nonmissing value for the nonmissing value X_{ij} of variable X_i. For the j^{th} observation, p_i is the number of variables that are nonmissing. Sites will be deleted if at least 3 observations out of the 4 index variables are missing.

After this, we have 2250 sites left which is about half of the original observations. We used the data and performed a simple imputation to fill the missing values for the observations that remain. Then we Box-Cox transformed stressors to make the distribution of each stressor more symmetric. We deleted the outliers with metal>5.7, anion>15 or nutri>8, which resulted in the 2211 observations to be used for our final hotspot detection. Table 5.4 is the summary of the data used for multivariate spatial partitions and the transformation used for each variable. Figure 5.2 is the distribution of reference sites and non-reference site in West Virginia.

Table 5.4 The summary of variables for the WV dataset in year 1996-2005.

Variable	N	Mean	Std Dev	Minimum	Maximum	Transformation
SAMPLE_LON_DD	2211	-80.641	1.074	-82.634	-77.753	
SAMPLE_LAT_DD	2211	38.669	0.707	37.217	40.616	
P_2Dom	2211	0.901	0.181	0.526	1.570	arcsin
P_chiro	2211	0.478	0.255	0.000	1.570	arcsin
P_EPT	2211	0.821	0.284	0.000	1.467	arcsin
HBI	2211	4.271	1.028	0.520	9.120	
EPT_Taxa	2211	7.654	3.726	0.000	18.000	
Total_Taxa	2211	15.035	4.607	1.000	29.000	
WVSCI	2211	67.193	17.628	9.780	98.550	
Cover	2211	14.423	3.375	2.000	20.000	
Embed	2211	12.906	3.509	0.000	20.000	
Velocity	2211	12.370	3.226	2.000	20.000	
Alter	2211	15.541	3.507	0.000	20.000	
Sediment	2211	12.136	3.753	0.000	20.000	
Riffl_Sinu	2211	15.694	3.585	0.000	20.000	
Chanflow	2211	14.241	3.437	2.000	20.000	
BnkStbTot	2211	14.241	3.538	1.000	20.000	
BnkVegTot	2211	13.545	3.920	0.000	20.000	
Temp	2211	18.191	4.131	6.980	30.610	
pH	2211	7.322	0.907	2.700	10.479	
DO	2211	8.901	1.428	1.020	19.399	
Conduct	2211	5.343	1.099	0.292	7.934	log
nutri	2211	0.058	0.920	-4.038	7.733	
ak_hard	2211	2.870	0.104	2.611	3.511	log
metal	2211	2.361	0.002	2.349	2.387	log(log)
anion	2211	1.972	0.466	0.713	4.163	log(log+3)

Figure 5.2 The distribution of test sites (in red) vs. reference sites (in green). Sites with WVSCI>78 are classified as reference sites and sites with WVSCI<78 are classified as test sites



5.5.3 RMPM result

By using the proposed refined multivariate partition modeling approach, we want to find the underlying multivariate structure over the region of West Virginia. After 100,000 simulation runs for 2 to 7 cluster partitions, we got the optimal weighed BIC values, $BIC_{W(k)}$, the $Diff_{WBIC_k}$ value, and the $Ratio_k$. $Ratio_k$ is the ratio of the change of the weighted BIC relative to the benchmark BIC value. Table 5.5 shows all these values for different partitions. Figure 5.3a is the plot of the optimal weighted BIC value vs. the number of partitions. Figure 5.3b is the scree plot of $Diff_{WBIC_k}$ vs. the number of partitions. Combining the scree plot and the $Ratio_k$ value, we will choose the *5-cluster* partition modeling as our optimal global partition model.

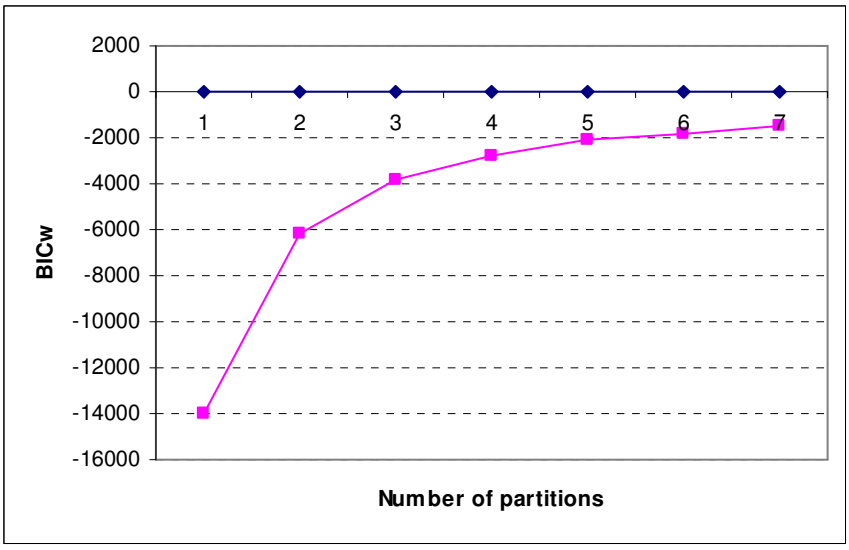
The *5*-cluster partition result for West Virginia is shown in Figure 5.4. Here the black area is the 1st cluster, the blue area is the 2nd cluster, the green the 3rd , the red the 4th, and the yellow is the 5th cluster. The first cluster is the natural area of the Potomac valley. The third cluster includes most of the Allegheny-Monongahela basin within West Virginia. The 4th and 5th clusters together form the majority of the Kanawha-New River Basin. Historically, there are two regions in West Virginia that have the most mining

activities. One is the northeast region, the other is the southwest region. The most productive coal mining counties in the northeast region are Monogalia, Marion, Preston, and Marshall counties and the most productive coal mining counties in the southwest region are Mcdowell, Wyoming, Logan, Boone, Raleigh, Kanawha and Mingo (see http://www.wvminesafety.org/PDFs/reserves2004.pdf). Figure 5.4 shows that the 3rd cluster captures the northeastern region and the 5th cluster captures the southwest region.

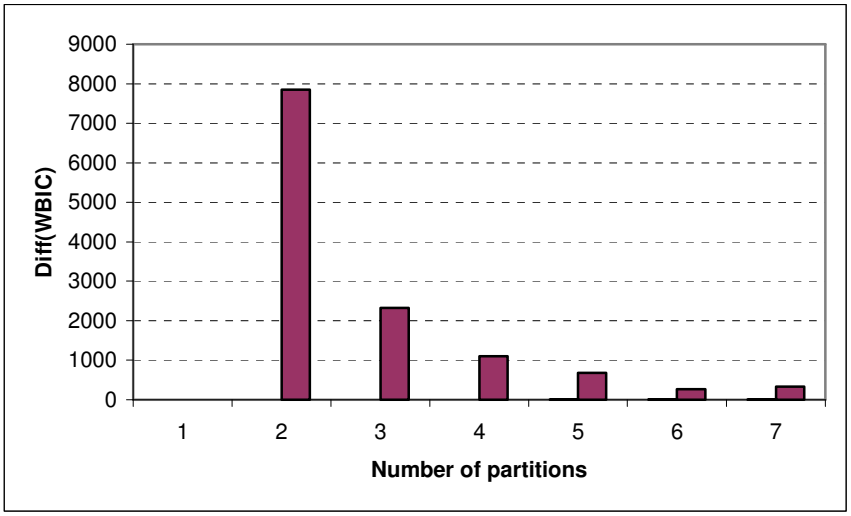
Table 5.5 Refined multivariate spatial partition modeling results of $BIC_{W(k)}$, $Diff_{WBIC_k}$ and $Ratio_k$.

Number of partitions	$BIC_{W(k)}$	$Diff_{WBIC_k}$	$Ratio_k$
1	-14021	--	--
2	-6169.26	7852.25	0.56002
3	-3847.59	2321.59	0.16557
4	-2745.41	1102.19	0.07861
5	-2067.73	677.71	0.04833
6	-1800.12	267.61	0.01909
7	-1468.32	331.80	0.02368

Figure 5.3 (a) The optimal weighted BIC (BICW) value vs. the number of partitions. (b) The plot of the difference in adjacent BICW values ($Diff_{WBIC_k}$) vs. the number of partitions.

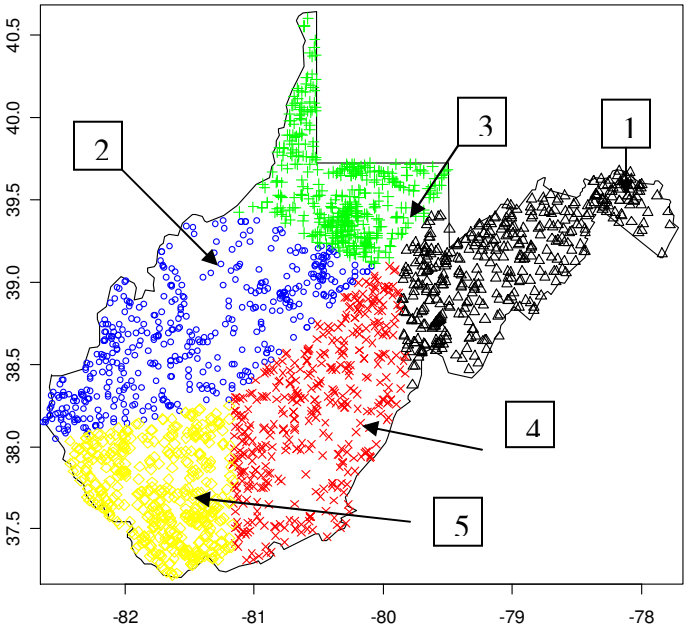


(a)



(b)

Figure 5.3 The 5-cluster optimal partition of West Virginia for years 1996-2005.



The RDA models for optimal 5-cluster partition are described in Table 5.6. The variable 'Conduct' appears in all RDA models and PH appears in the most models. The variable 'metal' only appears in the models of cluster 3, 4 and 5. The remaining variables

in the models are habitat assessment variables such as sand sedimentation, channel sinuosity etc.

Table 5.6 The RDA models for optimal 5-cluster partition for the WV dataset of year 1996-2005.

```
model for cluster 1 is:
rda(formula = y ~ Conduct + pH + ak_hard + Temp + Embed + Chanflow,
    data = x, scale = T)

model for cluster 2 is:
rda(formula = y ~ Conduct + Riffl_Sinu + Cover + Temp + pH + Velocity,
    data = x, scale = T)

model for cluster 3 is:
rda(formula = y ~ Conduct + pH + BnkVegTot + metal + Temp + Embed,
    data = x, scale = T)

model for cluster 4 is:
rda(formula = y ~ Conduct + metal + Embed + pH + Alter + Riffl_Sinu,
    data = x, scale = T)

model for cluster 5 is:
rda(formula = y ~ Conduct + BnkVegTot + Cover + metal + Riffl_Sinu +
    Chanflow, data = x, scale = T)
```

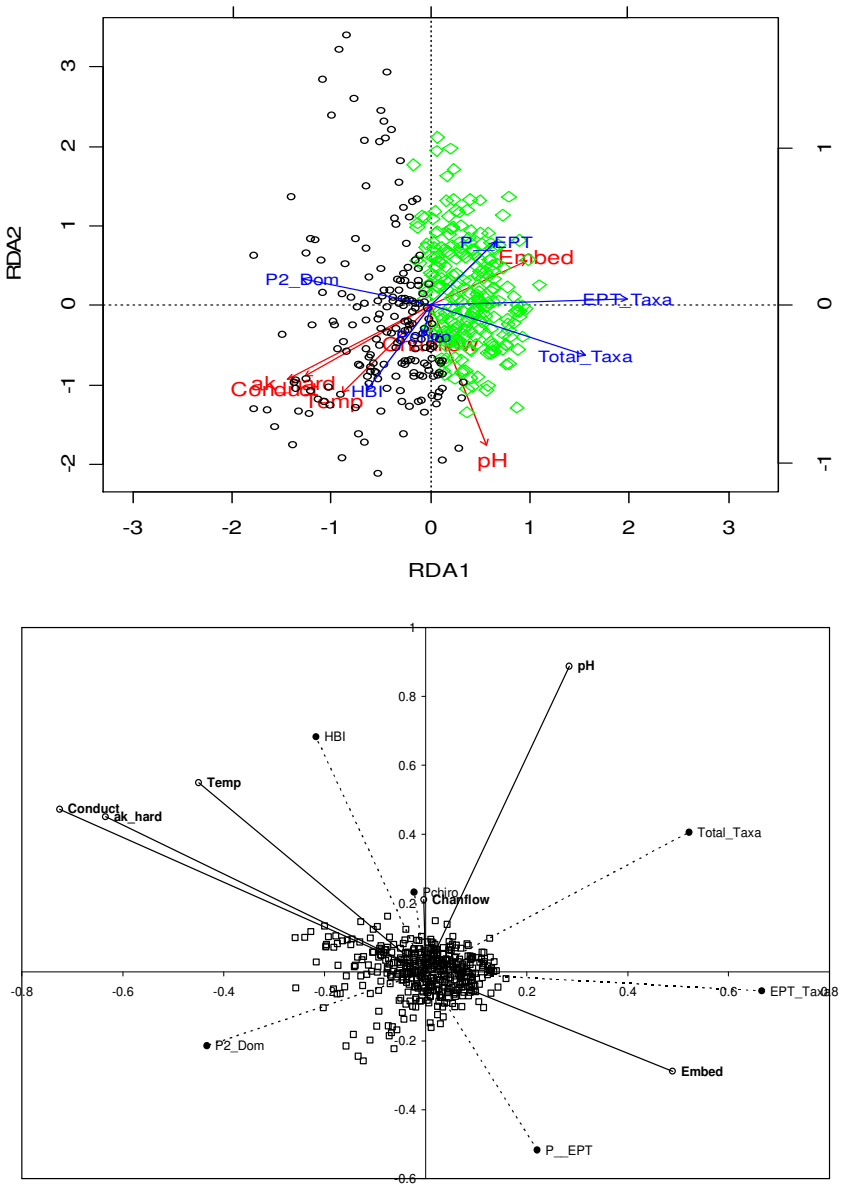
The triplots of the redundancy analysis for each cluster are shown in Figure 5.5 (a) to (e) separately. Figure 5.5a is the triplot for the first cluster. The region of this cluster is the Potomac basin. More reference sites exist than test sites. The EPA_Taxa ($r^2 = 0.48$) and Total_Taxa ($r^2 = 0.35$) are the two most sensitive biological responses to the stressors selected in this region. The stressor that has a significantly negative effect on these two responses is conductivity. Since this region is not a coal mining active area, we do not observal any coal mining effect in the triplot. Figure 5.5b is the triplot for the second cluster. In this region EPT_Taxa is the only response that is sensitive to the stressors selected ($r^2 = 0.38$). Temperature and conductivity have a negative effect on EPT_Taxa. Although this region has more test sites, they are not affected by mining activities since we can not find a mining effect in the triplot. Historically, this region is not an active mining production area either. Triplots for cluster 3 and 4 show that Conduct, metal and Embed have an effect on the biological responses for these regions. Figure 5.5c is the triplot for cluster 3. This cluster is located in the area of Allegheny-Monongahela basin within West Virginia. In this region, metrics EPT_Taxa, Total_Taxa, HBI and P_EPT are sensitive to the selected stressors: Conduct, metal, Embed, PH,

BnkVegTot and Temp. Metal has a major effect on the metric EPT_Taxa and P_EPT in this region. Due to previous active mining activity, most sites in this region are not in good condition. Figure 5.5d is the triplot for cluster 4. This cluster is located in the southeast part of the Kanawha-New River Basin. More than half of the sites in this cluster are reference sites. They are in very good condition. Although coal mining is not as active as in cluster 3, we detected mining effects in this region. From the triplot, we can see that Conduct, metal and Embed do work together to affect the biological metrics EPT_Taxa, HBI and P_2Dom. Resource managers should pay special attention to this area since the sites in good condition in this area may be at risk. Figure 5.5e is the triplot for cluster 5. This cluster includes the counties that have the most production of coal mining. Most sites in this region are in bad condition according to the WVSCI value because of the past mining activity. EPT_Taxa and Total_Taxa are sensitive to the stressors selected for this region. Among the stressors, Conduct and metal have the strongest effects on EPT_Taxa and Total_Taxa.

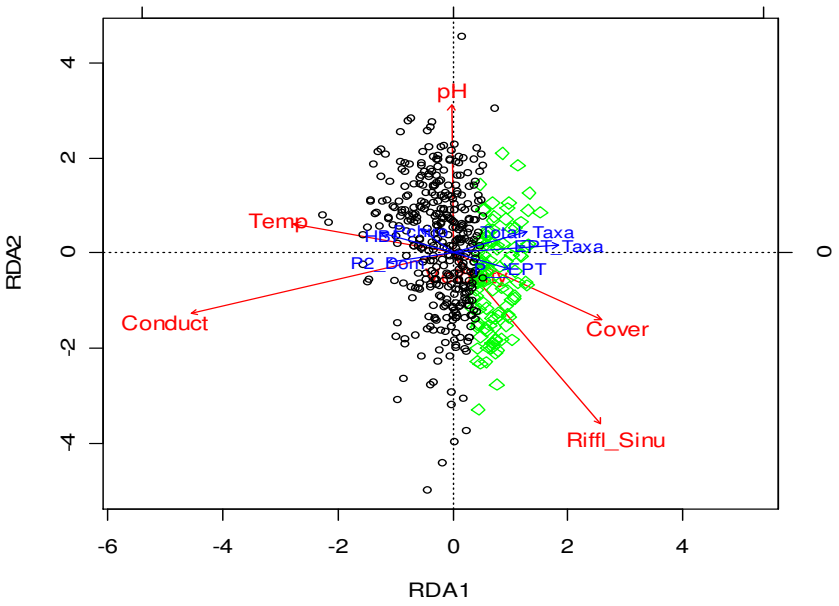
5.6 Concluding remarks

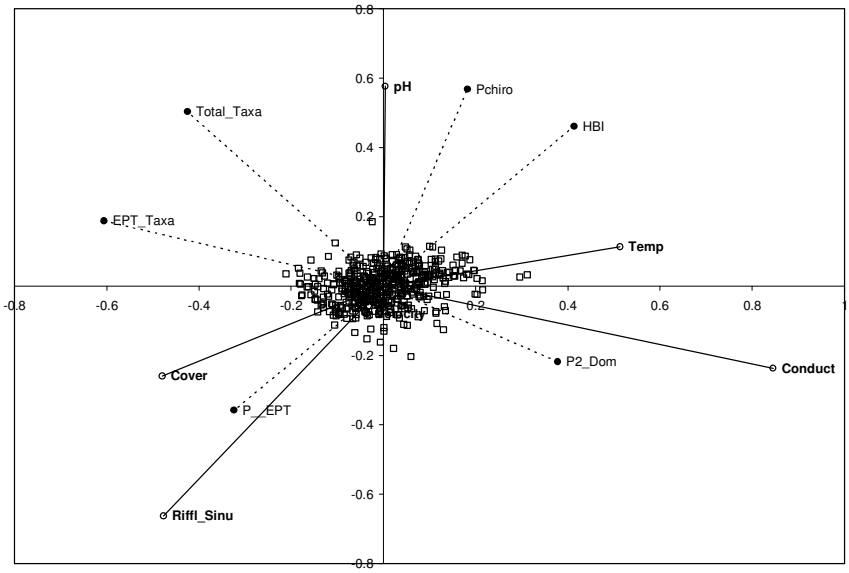
The multivariate spatial partition modeling approach provides us with a very useful tool to model multivariate response-stressor relationships over a large spatial region. By incorporating variable selection into the procedure, the proposed RMSPM has more flexibility to find the underlying relationship within each cluster. The application of this method to the West Virginia data in the previous section verified the appropriateness of the weighted BIC model selection criterion we proposed. The optimal partition we found is consistent with the historical facts about mining activities in West Virginia and the geographic regions.

Figure 5.4 Triplots for the 5-cluster partition. (a)-(e) Triplots for the first cluster to the fifth cluster.

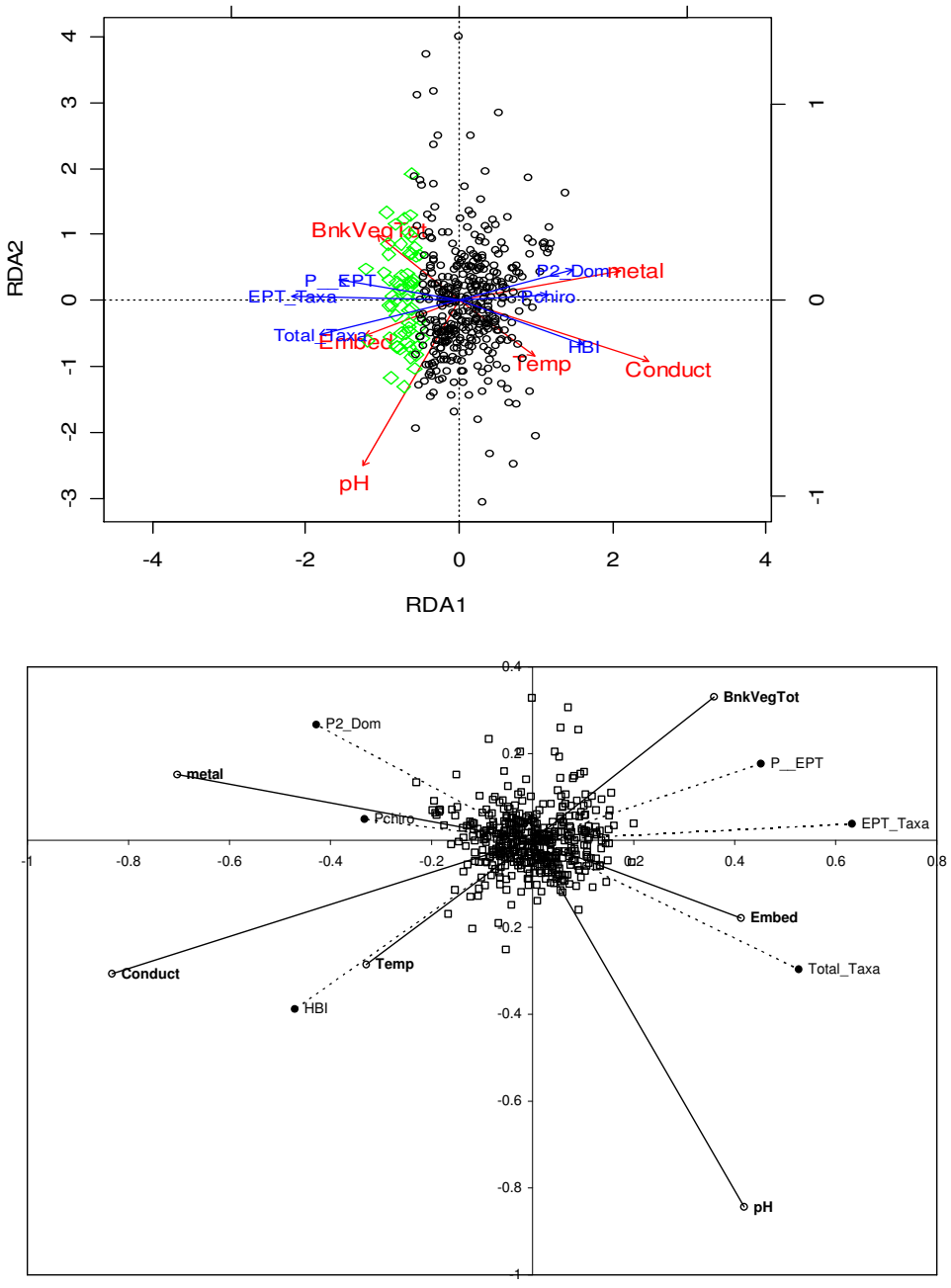


(a) Triplot for the first cluster.
Top plot:column scaling
Bottom plot: row scaling

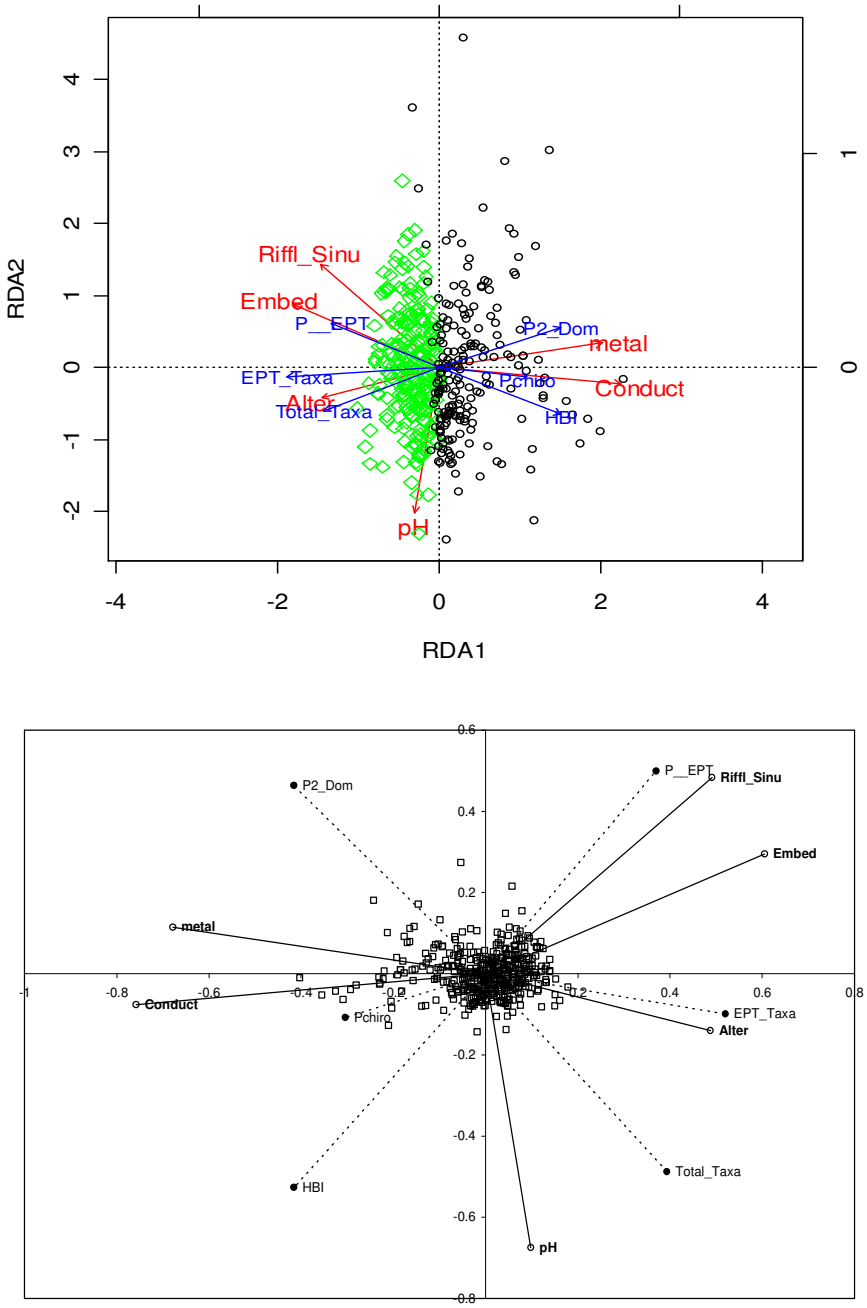




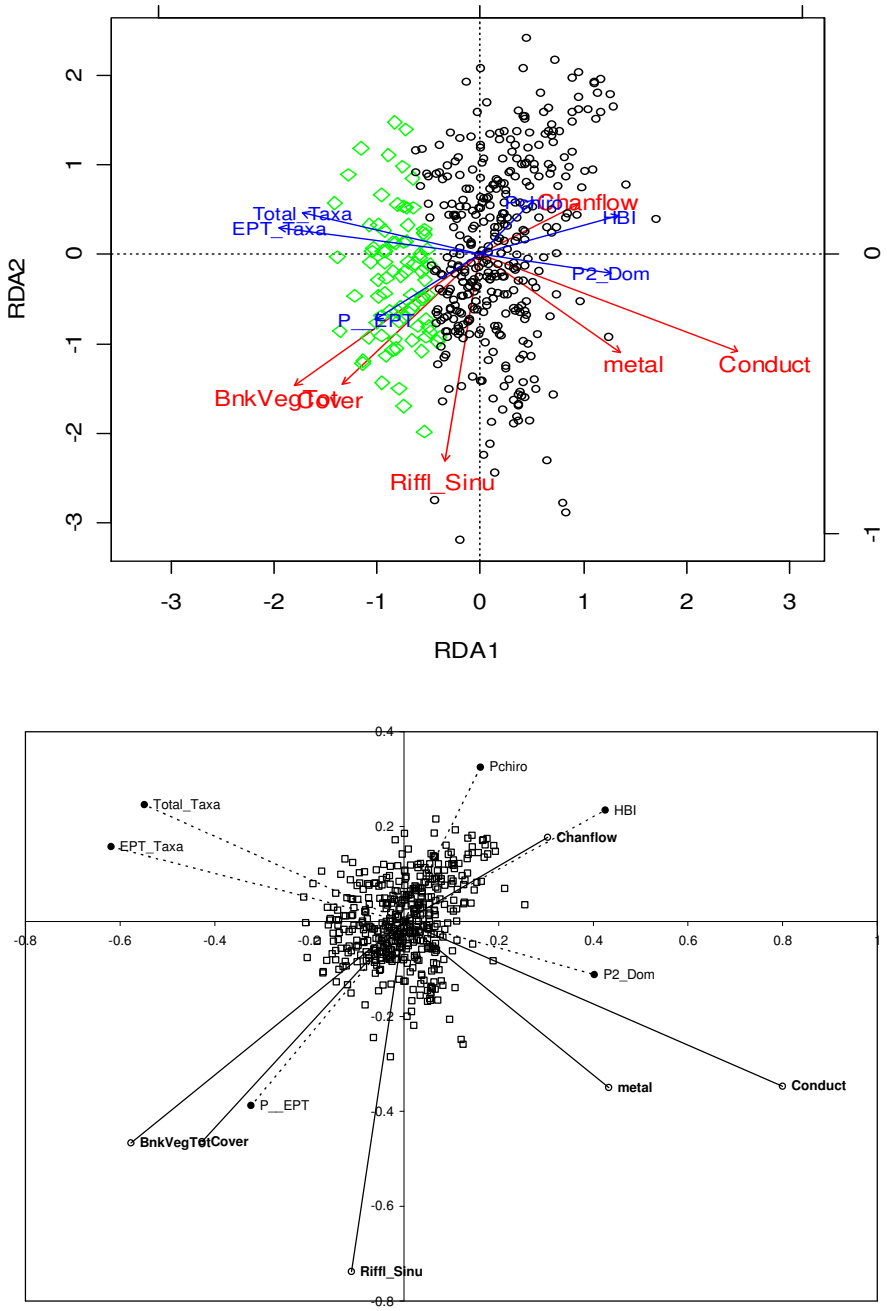
(b) Triplot for the second cluster.
Top plot: column scaling
Bottom plot: row scaling



(c) Triplot for the third cluster.
Top plot: column scaling
Bottom plot: row scaling



(d) Triplot for the fourth cluster.
Top plot: column scaling
Bottom plot: row scaling

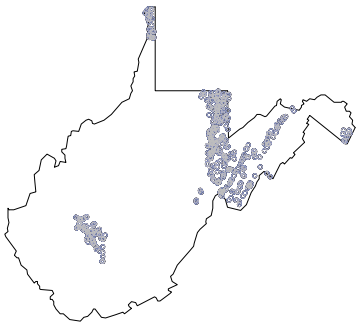


(e) Triplot for the fifth cluster.
Top plot: column scaling
Bottom plot: row scaling

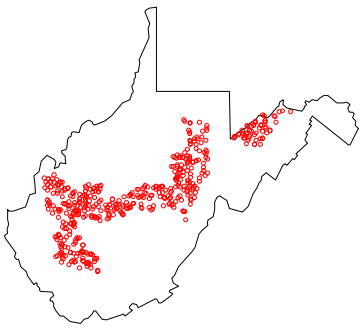
Appendix

Figure A5.1 The distribution of samples of West Virginia data by years.

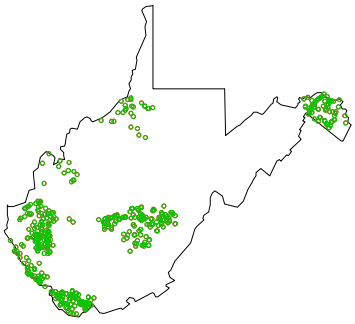
1996 (number of n=455)



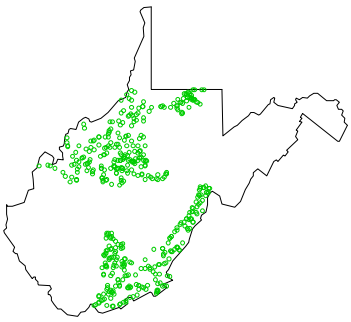
1997 (number of n=499)



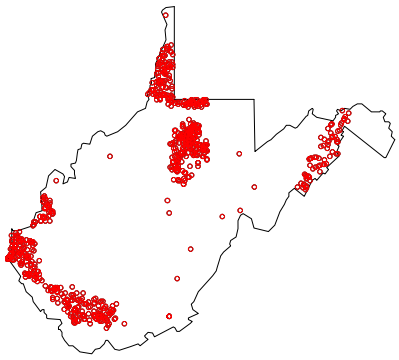
1998 (number of n=472)



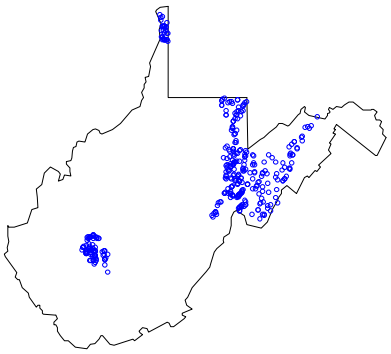
1999 (number of n=462)



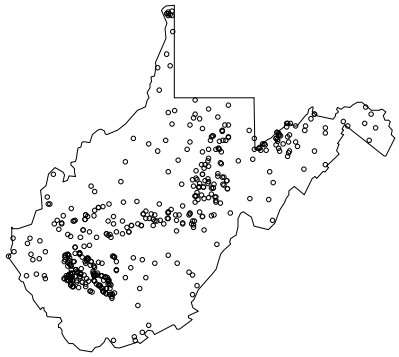
2000 (number of n=748)



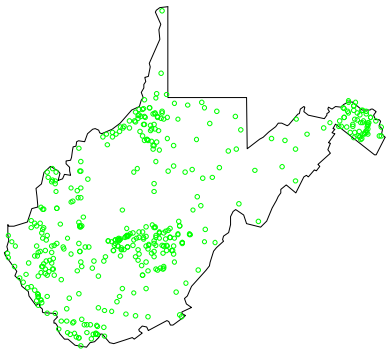
2001 (number of n=374)



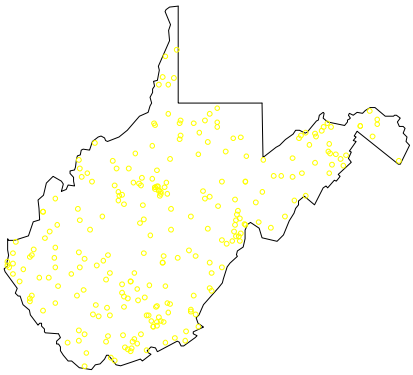
2002 (number of n=464)



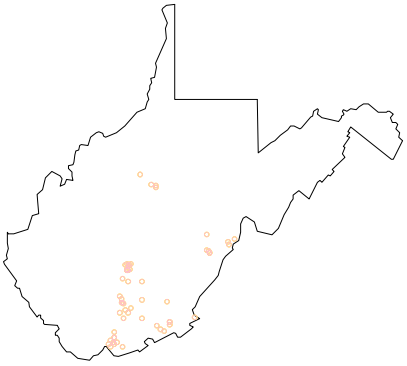
2003 (number of n=435)



2004 (number of n=252)



2005 (number of n=55)



WV samples for years 1996-2005 (number of n=4216)

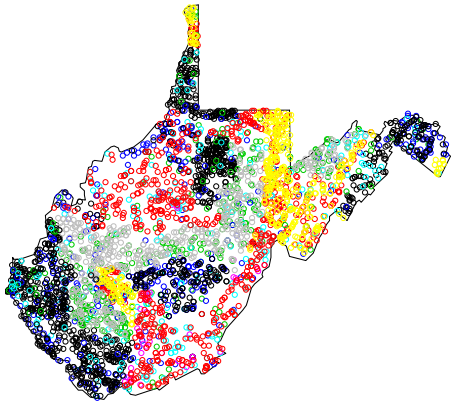


Table A5.1 The summary of variables of the WV dataset for years 1996-2005.

Variable	N	Mean	Type	Variable Label
Sample_ID	4216	6194.8700	6923.9900	
LON_DD	4216	-80.6419	1.0744	Longitude
LAT_DD	4216	38.6691	0.7078	Latitude
P_2Dom	4216	60.7142	15.9079	Percentage of 2 dominant Taxa
P_Chiro	4216	22.9592	20.7734	Percentage of Chironomidae
P_EPT	4216	54.8740	25.2024	Percentage of Ephemeroptera, Plecoprtera & Trichoptera Taxa
HBI	4216	4.3156	1.0408	Hilsenholf Biotic Index
EPT_Taxa	4216	7.5256	3.5648	Count of EPT taxa
Total_Taxa	4216	14.6816	4.4449	Count of total taxa
WVSCI	4216	67.0215	17.4836	Integrated biotic index
Cover	4208	14.2319	3.4044	Habitat vegetable/plant coverage
Embed	4205	12.8585	3.6266	Embededness
Velocity	4208	12.3659	3.2452	Stream velocity
Alter	4209	15.1553	3.5372	Stream Alterativeness
Sediment	4209	12.0368	3.8671	Sand sedimentations
Riffl_Sinu	4208	15.7043	3.5580	Channel sinuosity
Chanflow	4208	14.1314	3.5272	Tendency of stream to move back & forth
BnkStbTot	4208	14.0967	3.5444	Stream Bed Vegetation
BnkVegTot	4207	13.3617	3.8508	Bank total vegetation
RipVegTot	4207	10.3513	5.7500	Riparian vegetation
Total	4203	134.3071	23.4807	Total habitat assessment score
Temp	4016	18.2244	3.9911	Temperature
pH	3997	7.3754	0.8751	Dissolved Oxigen
DO	3984	8.8621	1.4075	Dissolved Oxygen
Conduct	4012	362.0282	523.0344	Conductivity
Fecal	3965	1471	8201.5000	
Acid_Hot	2252	7.8749	54.9057	
Alkalinity	2260	68.9801	69.0302	
Hardness	1907	141.0688	200.7721	
Sulfate	2250	139.1069	347.5572	
Chloride	1774	10.5732	27.9304	
Tot_Phos	2007	0.0465	0.1065	
NO2_NO3_N	2013	0.4063	1.2280	
Al_Tot	2355	0.7630	3.9655	
Ca_Tot	1930	37.8425	54.2881	
Cu_Tot	1709	0.0075	0.0675	
Fe_Tot	2329	1.2644	11.4277	
Mg_Tot	1910	11.0721	18.0258	
Mn_Tot	2321	0.2792	1.1612	
Zn_Tot	1799	0.0279	0.0811	

6 Searching for Multivariate Response-stressor Relationships

6.1 Introduction

During the last 20 years, many methods for finding a hotspot have been developed by researchers in different scientific fields. A hotspot can mean an unusual phenomenon, anomaly, aberration, outbreak, elevated cluster, or critical area (Patil & Taillie, 2003). Hotspot detection analysis is critically needed for monitoring, management, or early warning in a wide range of areas such as water resources, ecosystem health, public health and homeland security etc. Here we are interested in spatial hotspot detection in an ecological study to detect a spatially critical area where the biological community is threatened by degrading conditions caused by human perturbation.

Many of the spatial hotspot detection procedures use the basic idea of computing the number of cases within a circle or other geometric shape and then testing the count that results for statistical significance using a Poisson or Bernoulli type statistical test (McCullagh, 2006). Spatial scan statistics are used for these tests for spatial randomness, adjusting for the uneven geographical population density. McCullagh (2006) gave a good review on detecting hotspots in time and space. Although most methods used for hotspot detection are mathematically complicated, they are limited to finding the critical area based on the responses and their distributions, for example, the region where a crime rate is very high or the region where a disease population is significant. Generally, the response of interest is univariate.

In ecological studies, multivariate responses in the form of species counts or percentages are common and a univariate response usually can not explain whole community variation. In our study, we are interested in detecting the spatial region where the multivariate response-stressor relationship is the strongest, e.g. the region where coal mining activities have the strongest effect on biological species. The Voronoi based Multivariate Partition Method (MPM) introduced in a previous chapter will help us

partition the region of interest into clusters with different multivariate response-stressor relationships. In this chapter, we refine the method and propose an R-square-like criterion to find the area which has the strongest multivariate response-stressor relationship. We also incorporate the variable selection procedure for multivariate analysis into the MPM process. This resulting hotspot detection method is called the multivariate hotspot detection modeling approach (MHDM). Variable selection for multivariate regression can be carried out using a permutation test or an AIC-like criterion. An AIC-like measure can be used to build the model within clusters for an RDA/CCA analysis just as in the univariate variable selection procedure. The permutation test is used to judge the significance of a response-stressor relationship for one variable or for all variables used. In a typical environmental study, many uncertain factors exist and it is very important to incorporate a variable selection procedure into the entire clustering procedure so that the strongest relationship can be found in the region of interest. The AIC-like criterion is found to be consistent with the R-square-like model selection criterion. In this chapter, we will introduce the R-square-like hotspot detection criterion first; and then the permutation test and AIC-like criterion for variable selection in a multivariate direct ordination analysis will be described. The MHDM approach was applied to data from West Virginia in years 1995-2005, and also to the data in 2004 alone in order to find the region which has the strongest mining effect on biological responses. Note that the method assumes that a hotspot exists; a separate question is whether or not the detected hotspot is real.

6.2 R-square-like hotspot detection criterion

6.2.1 R-square-like quantities for a RDA analysis

Let $\mathbf{Y}$ be the response matrix and $\hat{\mathbf{Y}}$ be the fitted response matrix in the redundancy analysis. Redundancy analysis usually does not completely explain the variation in the response variables. The variation in the response variables explained by the environmental variables can be accounted by the sum of eigenvalues of the covariance matrix $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$ corresponding to the fitted values $\hat{\mathbf{Y}}$. So the proportion of

variation of biological community explained by the environmental variable is an approximation of a R-square-like quantity

$$r^2_{RDA} = \frac{\sum_{j=1}^{mc} \lambda_j^*}{\sum_{j=1}^{mc} \lambda_j},$$

where *mc* is the maximum number of the non-zero eigenvalues (see table 5.2 for detail), the λ_j^*s are non-zero canonical eigenvalues, the eigenvalues of the covariance matrix corresponding to the fitted values $\hat{\mathbf{Y}}$. The sum of these canonical eigenvalues $\sum_{j=1}^{mc} \lambda_j^*$ is the variance explained by the environmental variables. The λ_js are the non-zero eigenvalues of the covariance matrix corresponding to community data **Y**, called the unconstrained variance. The difference between an observed value **Y** and fitted value $\hat{\mathbf{Y}}$ is the residuals matrix $\mathbf{Y}_{res}$. $\mathbf{Y}_{res}$ has size $n \times p$ and the eigenvalues from this matrix are called non-canonical eigenvalues. Here the total variance of community variables **Y,** namely $\sum_{j=1}^{mc} \lambda_j$, can be expressed as the sum of the canonical eigenvalues and non-canonical eigenvalues.

6.2.2 R-square-like quantity for a CCA analysis

Essentially, a CCA is a weighted form of redundancy analysis, applied to the transformed matrix $\bar{\mathbf{Q}}$ (see 5.2.2). A CCA does not completely explain the variation in the response variables (matrix $\bar{\mathbf{Q}}$). The total variation in the response variables explained by the environmental variables can be accounted for by the trace of the variance covariance matrix corresponding to the fitted values of $\hat{\mathbf{Y}}$, i.e. $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$. So the proportion of variation in the response variable explained by the environmental variable is an approximation of the R-square-like quantity

$$r_{CCA}{}^2 = \frac{\sum_{j=1}^{mc} \lambda_j^{**}}{\sum_{j=1}^{mc} \lambda_j},$$

where *mc* is the maximum number of the non-zero eigenvalues, λ_j^{**} s are non-zero canonical eigenvalues, the eigenvalues corresponding to the covariance matrix of fitted values $\hat{\mathbf{Y}}$. The sum of these canonical eigenvalues $\sum_{j=1}^{mc} \lambda_j^{**}$ is the variance explained by the environmental variables, called the constrained inertia. The λ_j s are the eigenvalues of the covariance matrix of transformed community data $\bar{\mathbf{Q}}$, called non-canonical eigenvalues. The sum of these non-canonical eigenvalues is the variance explained by the community variables, called unconstrained inertia. Like a RDA analysis, the inertia to transformed community data $\bar{\mathbf{Q}}$, namely $\sum_{j=1}^{mc} \lambda_j$, can be expressed as the sum of canonical eigenvalues and non-canonical eigenvalues.

6.2.3 Hot spot detection criterion

In this chapter, our research interest is to find the specific spatial area within the whole region where the association of the response-stressor is strongest so that effects such as that of a mining can be detected. The optimal criterion will select the cell with the maximum R-square-like measurement.

In multivariate regression, we have $\mathbf{R_p^2 = (Y'Y - n\bar{y}\bar{y}')^{-1}(\hat{B}_p' X_p' Y - n\bar{y}\bar{y}')}$ to express the R-square measurement for fixed *p* regressors. This quantity can be approximated by our r_{RDA}^2 and r_{CCA}^2 for a given number of clusters. Let k be the index of the number of clusters, then R-square-like criterion for hotspot detection is

$\underset{k}{Max}\ r_{RDA}^2$ or $\underset{k}{Max}\ r_{CCA}^2$

Using this criterion, we usually try to select a small number of clusters eg, $k = 2,3,4,5$ to find the region of interest that contains a reasonably large number of observations and makes sense in our analysis.

6.3 Variable selection to get the maximum of the R-square-like quantity

6.3.1 An AIC-like criteria to build models within clusters

The Akaike Information Criterion (AIC) (Akaike, 1973) for model selection has the form $AIC = -2\ln[L(\hat{\theta})] + 2p$, where p is the number of free parameters in the model and $L(\hat{\theta})$ is the maximized log-likelihood. For standard linear models with residual sum square RSS, we can use $\log[L(\hat{\theta})] = -\frac{1}{2} n \ln(\hat{\sigma}^2)$, where n is the sample size and $\hat{\sigma}^2 = RSS/n$ (Burnham & Anderson, 1998).

Multivariate constrained ordination methods such as a RDA/CCA do not have a log-likelihood, which means that they cannot have an AIC and deviance (Okanen, 2006). We need to find a quantity which can give us a multivariate measure of AIC so that the model selection in multivariate analysis can be done in a reasonable way. Godinez-Dominguez and Freire (2003) use an analog of the univariate RSS for model selection in multivariate CCA, where $RSS =$ (sum of all eigenvalues)-(sum of the fitted eigenvalues). In our notation, this will be $RSS = (1 - r^2_{RDA} \text{ or } r^2_{CCA}) * \sum \lambda$. Using this RSS, $AIC = n \ln RSS + 2p + \text{constant}$ is our model building tool.

6.3.2 Validation of using an AIC-like variable selection criterion for hotspot searching

Our research interest here is to find the hotspot area where the association of response to stressors is the strongest within the whole region. The optimality criterion in the multivariate spatial partition modeling approach for hotspot detection will select the

spatial area with maximum r^2_{RDA} or r^2_{CCA}. The AIC-like variable selection criterion matches this R-square-like model selection criterion very well in that it will select the variables (model) within each cluster that minimize the AIC-like quantity, minimize the RSS, or maximize the R-square-like measurement within each cluster. This will give us a much better way to find the hotspot and the model within the hotspot than the method where only a fixed number of stressors are used to build a model within the hotspot.

6.3.3 Permutation test for significance of the response-stressor relationship

In ecological studies, it is common to evaluate the statistical significance of the RDA response-stressor relationship using a permutation test. The null hypothesis is that the responses are not related to environmental stressors and the alternative hypothesis is that the responses vary with the stressors. The basic steps for a standard statistical test include five steps (ter Braak & Smilauer, 1998):

(1) Choose a test statistic, F, that measures the strength of response-stressor relationship.

(2) Use the data to calculate the test statistic and get the value F_0.

(3) Generate l new datasets that are equally likely under the null hypothesis by permuting the samples in the response (species) data.

(4) Calculate the test statistic for each new dataset and get $F_1, F_2, ..., F_l$.

(5) The significance level of this permutation test is determined by the rank of F_0, $R_{(F_0)}$ among all values F_0 to F_l. Specifically, it is $1-\frac{R_{(F_0)}}{l+1}$.

The mathematically derived reference distribution for regression and ANOVA is the F-distribution and holds true if the data are independent and follow a normal distribution with homogeneous variance. In contrast to the standard statistical test, the reference distribution of the permutation test is determined by the data itself without the assumptions of normality. Under the null hypothesis, samples in the response data set are randomly linked with samples of environmental data, each permutation of samples of the

response data leads to a new dataset from which we can calculate the test statistic that should not vary much from each other. The reference distribution is the distribution of the test statistic in the permuted dataset. If the original set of values is quite different, this would suggest a significant relationship.

The above permutation test procedure has a null hypothesis of exchangeability of rows of responses. In our study, we will use the residuals of a linear model such as the RDA model as our permutation units. So in this case, the null hypothesis is that the residuals of the response variables after fitting some explanatory variables using the linear model are exchangeable. The reference statistic F, for testing the significance of the relationship in a RDA analysis is (Legeredre, 1998)

$$F = \frac{\sum_{j=1}^{mc} \lambda^* / p}{(\sum_{j=1}^{mc} \lambda - \sum_{j=1}^{mc} \lambda^*) / (n - p - 1)}$$

where p is the number of stressors, n is the number of observations, mc is the maximum number of the non-zero eigenvalues and $\sum_{j=1}^{mc} \lambda - \sum_{j=1}^{mc} \lambda^*$ is the residual sum of squares. For a RDA/CCA analysis, the permutation test procedure is as following,

(1) Calculate the test statistic for the data and get F_0.

(2) Regress the response $\mathbf{Y}$ on $\mathbf{X}$ and find the fitted values $\hat{\mathbf{Y}}$ and residual $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$. Permute $\mathbf{R}$ to get $\mathbf{R}^*$ and calculate the new dataset $\mathbf{Y}^* = \hat{\mathbf{Y}} + \mathbf{R}^*$.

(3) Calculate the test statistic F for each new data set $\mathbf{Y}^*$.

Repeat steps (2) and (3) l times to get $F_1, F_2, ..., F_l$.

(4) The significance level of this permutation test is determined by the rank of F_0, $R_{(F_0)}$, among all values F_0 to F_l. Specifically, it is $1 - \frac{R_{(F_0)}}{l+1}$.

The above procedure is used to test the significance of the response-stressor relationship in the hotspot region detected.

6.4 Multivariate hotspot detection modeling approach (MHDM)

The multivariate hotspot detection modeling approach (MHDM), defined below, is designed to find the hotspot region and select the important variables relating the biological responses with stressors.

1. Partition the two dimensional spatial region of data into k non-overlapping clusters using the Voronoi diagram technique.
2. Use a RDA/CCA analysis within each cluster using the AIC-like criterion to build the model. Calculate the value of the R-square-like criterion for each cluster. Find the largest R-square-like value for this k-cluster partition.
3. Repeat steps 1 and 2 a sufficient number of times such that the optimal k cluster with the largest R-square-like quantity is found.
4. Repeat steps 1 through 3 for $k = 2$ to $Max_k = K$ and find the optimal R-square-like values for the partitions with 2 to K clusters. Max_k is decided by experience and the minimum number of observations for each cluster.
5. Find the hotspot by choosing the cluster which has the highest R-square-like value all the partitions.

An alternative approach to MHDM is to run the multivariate spatial partition modeling procedure using all environmental variables as stressor variables and to find the hotspot which has the highest overall response-stressor relationship. Then use a variable selection procedure based on a permutation test or an AIC approach to find the build the model within the hotspot. We call this method the after-partition modeling approach. This method can be used as a validation method to verify our findings in the above approach. The result is not as precise as the MHDM approach.

6.5 Application of MHDM to West Virginia data

The West Virginia data set includes 4216 samples of benthic macroinvertebrates sampled from 1996 to 2005. Section 5.5.1 provided the details on the response and stressor variables and their properties. Figure A5.1 gave the yearly distribution of the

samples. We can see that 2004 had the most evenly distributed observations over the entire state of West Virginia. It is our interest to use the dataset for this year to verify our findings. The data was manipulated for better analysis. Section 5.5.2 provided the details on how the data was manipulated. We have 2211 observations left after the manipulation. In this section, we use the same data set to detect the hotspot of mining effect in West Virginia.

6.5.1 The benchmark model result

This section will describe the result of a benchmark model of multivariate RDA analysis using the entire 2211 observations across West Virginia in 10 years. First we perform an RDA analysis with variable selection using an AIC-like criterion. The model selected is Responses~ Conduct + pH + BnkVegTot + Embed + metal + Riffl_Sinu . The responses here are the six SCI index variables. The triplot of the relationship is shown in Figure 6.1. The sites on the left side of the first RDA axis are reference sites with higher values of metrics of EPT_Taxa and Total_Taxa. Sites to the right are the poorer quality sites with higher values of metrics of HBI and P_2Dom. The reference sites tend to have higher values for stream embeddedness and bank vegetation and lower values for metals and conductivity. From table 6.1 we can see that two axes are good enough to express the multivariate response-stressor relationship, since 96.8% of the variance of the fitted values is represented by a two dimension biplot. The unconstrained variance is 1 and the constrained variance is the sum of the canonical eigenvalues. The R-square-like value is 0.3114. Table 6.2 indicates that the univariate relationships are weak and that only EPT_Taxa is sensitive to the stressors. Regression coefficients for EPT_Taxa confirm that the biological quality increases with embededness and bank vegetation and decreases with conductivity and heavy metals.

Figure 6.1 A triplot of RDA analysis for WV dataset in 1996-2005.

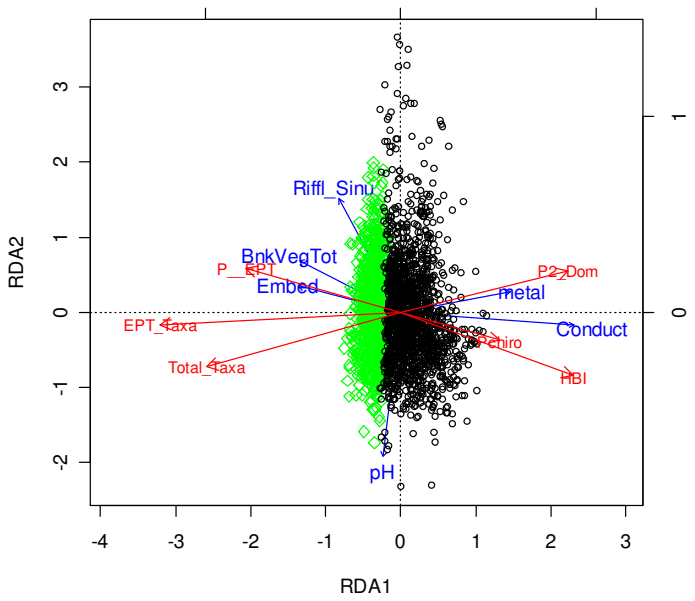


Table 6.1 The result of RDA analysis of WV dataset for years 1996-2005.

Statistics	Axis1	Axis2	Axis3	Axis4	Total variance
Eigenvalues	0.274	0.028	0.009	0.001	1.000
Species-environment correlations	0.703	0.374	0.233	0.131	
Cumulative percentage variance					
of species data	27.4	30.1	31.1	31.1	
of species-environment relation	87.9	96.8	99.7	99.9	
Sum of all eigenvalues					1.000
Sum of all canonical eigenvalues					0.3114

Table 6.2 Multiple regression and R-square values for each response in the benchmark model.

Regression coefficients: Transformed Y's are regressed on Transformed X's						
	P_2Dom	P_Chiro	P_EPT	HBI	EPT_Taxa	Total_Taxa
Embed	−0.0921	−0.0742	0.1267	−0.0264	**0.1234**	0.0887
Riffl_Sinu	−0.0024	−0.1062	0.1214	−0.1387	**0.0289**	−0.0606
BnkVegTot	−0.1104	−0.1307	0.09519	−0.1153	**0.1195**	0.1293
pH	−0.2681	−0.0157	−0.0314	0.0828	**0.1862**	0.2771
Conduct	0.4304	0.0837	−0.2457	0.3985	**−0.6223**	−0.4980
Metal	−0.0094	0.1516	−0.1779	0.1044	**−0.0439**	−0.0469
R squared:						
	0.2752	0.1169	0.2439	0.3188	**0.5381**	0.3749

6.5.2 Hotspot detection results

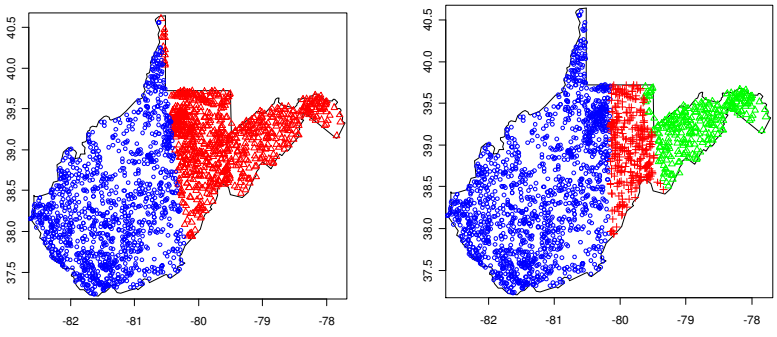
We first use the multivariate hotspot detection modeling approach to find the hotspot of the strongest multivariate responses-stressors relationship. Before the analysis, we do not know the model structure and we want to know the potential stressors that are highly related with the responses. We run 10,000 Voronoi tessellations for each cluster size ranging from 2 to 5 to help us find the hotspot and the model structure within. The number of samples within a cluster was restricted to have a minimum of 266 observations, which is 12% of the total observations. With the model selection procedure is incorporated inside the partition procedure, the multivariate hotspot detection modeling approach gives us maximum R-square-like values as well as the model structures within the hotspots detected for each of the 2-5 cluster partitions.

Table 6.3 summarizes the results of clustering. We find that cluster 4 in the 5-cluster partition is the global hotspot and the biological responses have the strongest linear relationship with conductivity, pH, heavy metals, bank plant coverage, sedimentation and dissolved substances. The global hotspot has 268 observations. Figure 6.2 maps the hotspot (red) regions and observations on the West Virginia map for 2- to 5-cluster partitions. The red region for the 5-cluster partition is the global optimal hotspot. The hotspots detected for the 3, 4 and 5-cluster partitions are approximately located on the northeast side of West Virginia. The models within the hotspots are slightly different.

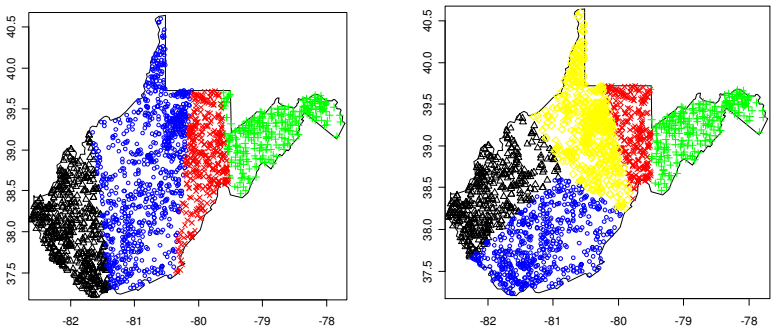
Table 6.3 Hotspot detection results using the MHDM modeling approach for 1995-2006 WV data.

Number of partitions	R-Square-like value /obs	Maximum R-Square-like	Model within
1	**0.3113/2211**	0.3113	**yy~ Conduct + pH + BnkVegTot + Embed + metal + Riffl_Sinu**
2	**0.4200/876**	0.4200	y ~ Conduct + pH + metal + BnkVegTot + Riffl_Sinu + Embed
	0.2498/1335		
3	0.2264/290	0.5107	y ~ Conduct + pH + metal + Embed + Temp + ak_hard
	0.2802/1572		
	0.5107/349		
4	0.2427/561	0.5221	y ~ Conduct + pH + metal + Embed + Temp + Alter
	0.3052/937		
	0.2254/316		
	0.5221/361		
5	0.2105/423	**0.5406**	**y ~ Conduct + pH + metal + BnkVegTot+ Sediment + anion**
	0.3142/648		
	0.2203/280		
	0.5406/268		
	0.3477/592		

Figure 6.2 The hotspots for the 2- to 5-cluster partitions (red area). The minimum number of observations per cluster is 266. The red area in (d) is the global hotspot detected.



(a) (b)



(c) (d)

6.5.3 RDA result(s) and discussion within the hotspot

Table 6.4 is the result of redundancy analysis for the West Virginia data sampled from 1996 to 2005 within the hotspot region. The sum of the canonical eigenvalues for the analysis is 0.541. The percentage of the variance explained by the first two axes was 95.8%. The variance of the biological community is 1. The R-square-like value improves from 0.311 in the benchmark RDA analysis to 0.541 in the hotspot RDA analysis. 999 permutation tests indicate that the model relationship within the hotspot region is significant (p=0.0002).

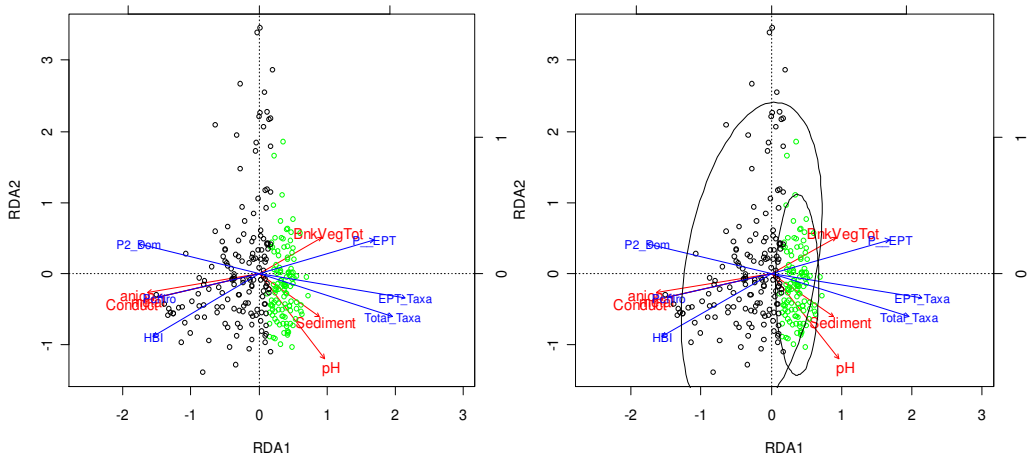
Table 6.4 Result of the redundancy analysis for 1996-2005 WV data within hotspot region.

Statistics	Axis1	Axis2	Axis3	Axis4	Total variance
Eigenvalues	0.474	0.044	0.022	0.001	1.000
Species-environment correlations	0.853	0.488	0.514	0.139	
Cumulative percentage variance					
of species data	47.4	51.8	53.9	54.0	
of species-environment relation	87.7	95.8	99.8	99.9	
Sum of all eigenvalues					1.000
Sum of all canonical eigenvalues					0.541
Test of significance of all canonical axes				P-value* =	0.0020
Test of significance of first canonical axis				P-value* =	0.0020

The P value is the result of 999 permutation tests

Figure 6.3 is the triplot for the RDA analysis in this hotspot region. Table 6.5 gives the correlation of the stressors in the model with the first 2 ordination axes. From the triplot we can see that the first axis is closely (negatively) related with the heavy metal, conductivity, and anion variable and positively correlated with the sediments and bank vegetable coverage (also see Table 6.5 for the corresponding signs). Those environmental variables correlated with the first RDA axis describe a gradient from sites with greater sediments and vegetable coverage at the positive end to the sites with greater heavy metal concentration and conductivity in the negative end. This axis is associated with mining effects. The mining effect is decreasing along the first axis from the triplot. The second axis is closely correlated with pH. The ellipses in Figure 6.3b corresponds to a 95% probability region for the test site scores (in black) and for the reference site scores (in green). We can also see that the reference sites have less mining effects than the test sites since all of the reference sites are located at the end with less heavy metals and conductivity on the first axis.

Figure 6.3 Triplot of RDA within hotspot region detected using MHDM approach. Black dots are the sites whose WVSCI <78 and green dots are the reference sites whose WVSCI >=78. (a) Triplot of RDA (b) Triplot of RDA with 95% probability ellipses for the reference and test groups.



(a) (b)

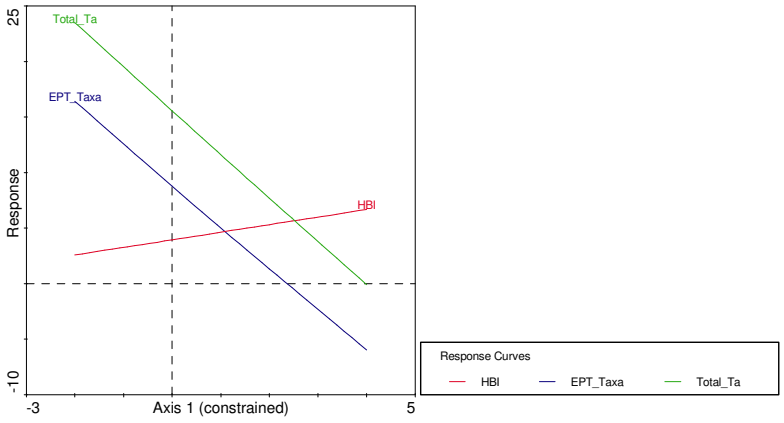
Figure 6.4 describes the estimated linear relationships between biotic metric (responses) and the first axis for the hotspot region. EPT_Taxa and Total_Taxa have negative linear relationships with this axis and HBI is positively related with it. In other

words, we can say that a high mining effect will increase HBI and decrease EPT_Taxa and Total_Taxa. P_2Dom and P_Chrio have positive relationships with the first axis and P_EPT is negatively related to it. In other words, we can say that high mining effects are highly correlated with the percentage of 2 dominant Taxa, the percentage of Chrionomidae and the percentage of EPT Taxa. All these findings are consistent with the expert findings of mining effects on the macroinvertebrate metrics (Sams III & Beer, 2000).

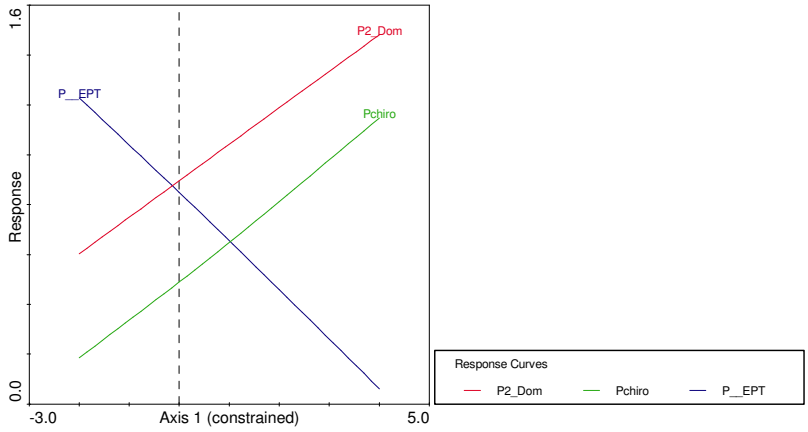
Table 6.5 Correlation of the environmental variables with the first two axes.

	Axis 1	Axis 2
Sediment	0.4601	-0.3161
BnkVegTot	0.4804	0.2646
pH	0.4989	**-0.6179**
Conduct	**-0.8837**	-0.1953
metal	**-0.7713**	-0.1699
anion	**-0.8471**	-0.1349

Figure 6.4 Plot of the estimated linear relationship with the first axis for (a) EPT_Taxa, Total_Taxa and HBI (b) P_2Dom, P_Chiro and P_EPT.



(a)



(b)

Table 6.6 describes how each biological metric (response) changes linearly with each stressor as illustrated in the two dimensional triplot in Figure 6.4. It also gives the R-square values for individual responses after the RDA analysis. All of the R-square values have huge improvements within the hotspot region. The R-square value for EPT_Taxa jumped from 0.5381 for the benchmark model to 0.7396 for the hotspot model. The R-square value for P_2Dom, P_Chiro, P_EPT, HBI and Total_Taxa increase from 0.2752, 0.1169, 0.2439, 0.3188 and 0.3749, respectively, in the benchmark model to 0.5015,0.4063, 0.4632, 0.5042 and 0.6287, respectively, in the hotspot model, an average increase of more than 20%. In the hotspot model, the majority of the responses have R-square values that are greater than 0.5.

Table 6.6 Linear relationships between species scores and environmental scores for the first two axes and R-square values for the individual responses.

Regression coefficients: Transformed Y's are regressed on Transformed X's						
	P_2Dom	P_Chiro	P_EPT	HBI	EPT_Taxa	Total_Taxa
Sediment	-0.1508	-0.0054	0.0117	0.0547	0.1496	0.1226
BnkVegTot	-0.1024	-0.0883	0.1150	-0.1496	0.1442	0.0926
pH	-0.3504	-0.0022	-0.0446	0.2261	0.3724	0.4864
Conduct	0.2025	-0.0185	-0.2612	0.4285	-0.6105	-0.3881
metal	0.0034	0.4938	-0.4067	0.2846	-0.1224	-0.0782
anion	0.2479	0.1872	-0.0792	0.0814	-0.0745	-0.1548
R squared:						
	0.50145	**0.40625**	**0.46320**	**0.50422**	**0.73966**	**0.62868**

6.5.4 Analysis using after-partition modeling approach

How do we know that the global hotspot that we detected and the model within the hotspot is the one closer to the real world model? This can be verified by an after partition modeling approach and checking with field managers or professionals. Table 6.7 is the result of hotspot detection using a multivariate partition modeling approach with all stressors. The third cluster in the 3-cluster solution has the highest R-square-like value of 0.5667 with 279 observations. The red area in Figure 6.5a is the location of this hotspot on the West Virginia map. Figure 6.5b is the hotspot detected using the multivariate hotspot detection modeling approach from the last section.

The AIC-like variable selection and the stepwise method by a permutation test select the same set of stressors within the hotspot region. This is shown in Table 6.8. It is not a surprise since the AIC-like criteria will choose variables that minimize the RSS which is consistent with the mechanics of the variable selection using the permutation test. The model is similar to the model from hotspot detection using the MHDM approach (Table 6.3) except that the variable anion is substituted by DO.

Table 6.7 R-square like value for hotspot detection partition using multivariate partition modeling approach with all stressors

Number of partitions	R-Square-like value within/obs	Maximum R-Square-like
1	**0.3114/2211**	0.3114
2	**0.4431/280**	0.4431
	0.3257/931	
3	0.3084/1666	**0.5666**
	0.2542/266	
	0.5666/279	
4	0.3241/1135	0.5489
	0.2998/429	
	0.2506/379	
	0.5489/268	
5	0.2698/283	0.5536
	0.5536/296	
	0.3384/354	
	0.2852/617	
	0.3029/661	

Table 6.8 Model building within the hotspot region using permutation tests and an AIC criterion

Variable selected by permutation test (permutation=999) by CANOCO

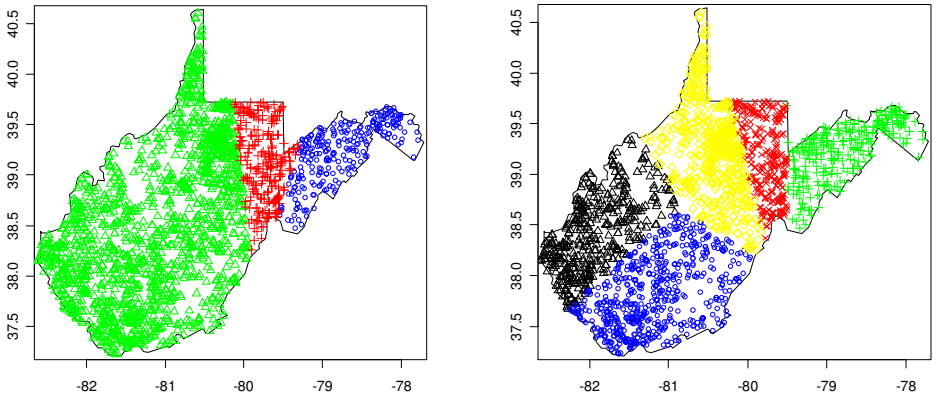
Variable	P	F
Conduct	0.001	143.53
pH	0.001	58.90
metal	0.001	29.96
Sediment	0.001	8.68
BnkVegTot	0.003	4.71
DO	0.010	3.92

Variables selected by using AIC-like criterion

Conduct, pH, metal, Sediment, BnkVegTot, DO

Table 6.9 shows that 94.5% of the variance of fitted responses can be explained by the stressor variables selected. The model selected is significant with p-value=0.002. Figure 6.6 is the triplot within the hotspot region. The interpretation for this plot is very similar to the one in Figure 6.3.

Figure 6.5 (a) The red area is the global hotspot region detected using after partition modeling approach (b) The red area is the hotspot detected using MHDM modeling approach. The minimum number of observations per cluster is 266.



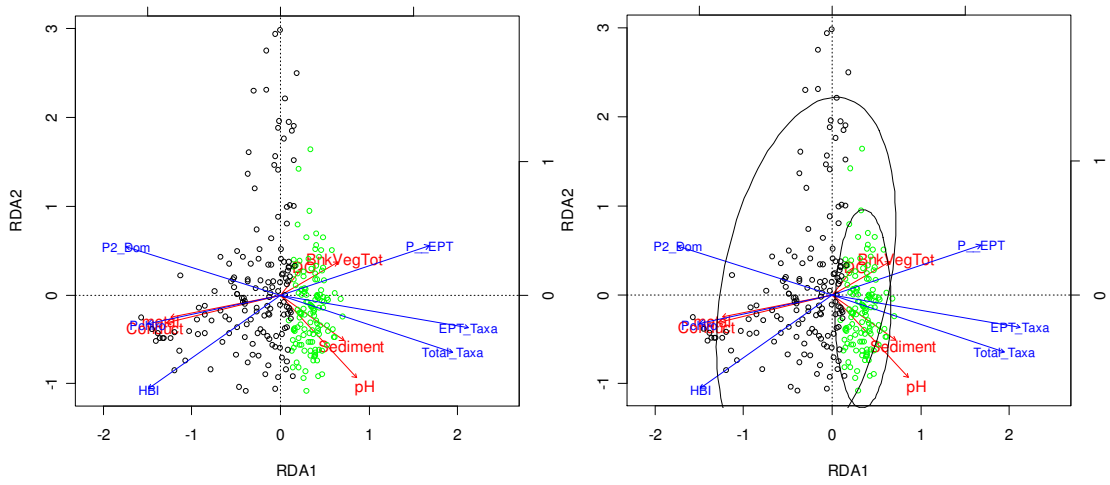
(a) (b)

Table 6.9 Result of the redundancy analysis within the hotspot region.

Statistics	Axis1	Axis2	Axis3	Axis4	Total variance
Eigenvalues	0.455	0.058	0.026	0.001	1.000
Species-environment correlations	0.852	0.537	0.535	0.204	
Cumulative percent variance of species-environment relation	84.2	94.9	99.7	100.0	
Sum of all eigenvalues					1.000
Sum of all canonical eigenvalues					0.566
Test of significance of all canonical axes			P-value*	=	0.0020
Test of significance of first canonical axis			P-value*	=	0.0020

* The P value is the result of 999 permutation tests

Figure 6.6 Triplots of RDA within the hotpot region detected using the after partition modeling approach. Black dots are the sites whose WVSCI <78 and blue dots are the reference sites whose WVSCI >=78. (a) A triplot of a RDA (b) A triplot of a RDA with 95% probability ellipses for the reference and the test groups.



6.6 Analysis of the West Virginia 2004 data

6.6.1 Data manipulation before analysis

As a second example, consider the 252 total sites sampled in West Virginia in 2004.Table 6.10 is the summary of the mean of the biotic indices, habitat bioassessments, field chemical/physical measurements and lab chemical/physical measurements. All the GIS variables such as elevation and watershed area have fewer than 10 observations and Cu_Tot & Zn_Tot had only 12, so none of these variables will be included in our analysis.

The habitat and chemistry measurements (stressors) have a significant amount of redundancy. To reduce the colinearity in the regression analysis, we selected a reduced environmental data set. The selection is done by the following steps. First, we evaluate the Pearson product-moment correlation among the seven SCI metrics with all the habitat and chemistry variables. The chemistry and habitat variables that had a correlation greater than 0.2 with at least one SCI index were kept for analysis. The 17 variables remaining were: Cover, Embed, Alter, Sediment, Riffl_Sinu, BnkVegTot, Temp, pH,

Conduct, Alkalinity, Hardness, Sulfate, Chloride, Al_Tot, Ca_Tot, Mg_Tot and Mn_Tot. The Pearson product-moment correlation among the habitat and chemistry variables was evaluated. We found that Embed, Cover, and Sediment had high pairwise correlations. Al_Tot had a high correlation with the heavy metal Fe (r=0.73) and Mn (r=0.52). Alkalinity, Hardness, Sulfate, Magnesium and Calcium are dissolved forms with high pairwise correlations (r>0.5). These will be useful for the later interpretation of the hotspot triplot. For all the variables, missing values were imputed and the Box-Cox transformation was used to reduce skewness.

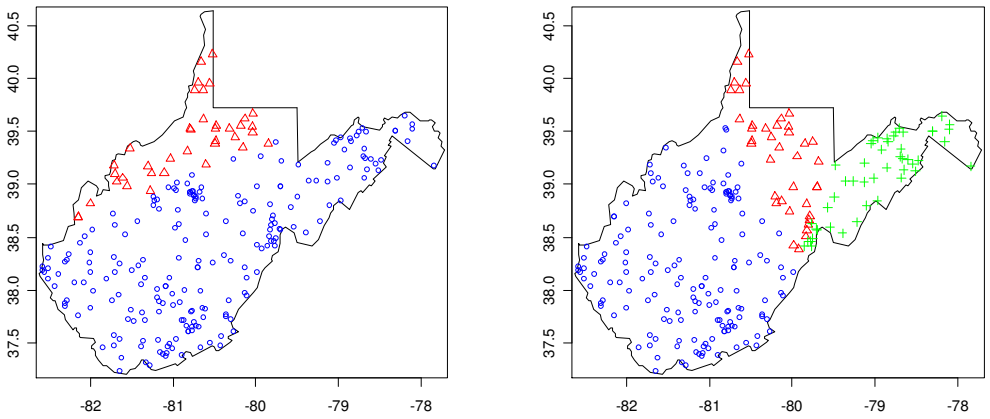
Table 6.10 Summary of the West Virginia 2004 data.

Variable	N	Mean	Std Dev	Minimum	Maximum
P_2Dom	252	58.3146	15.6782	29.72	98.0000
P_Chiro	252	21.8810	21.1124	0.4200	94.3300
P_EPT	252	62.8031	23.2743	0.9100	96.2000
HBI	252	3.9479	0.9785	0.7200	6.0400
EPT_Taxa	252	9.1547	3.4795	2.0000	18.0000
Total_Taxa	252	15.8055	4.2103	3.0000	27.0000
WVSCI	252	72.8382	17.1977	19.5000	97.4300
Cover	252	13.8055	3.1770	2.0000	19.0000
Embed	252	12.9206	3.7108	2.0000	20.0000
Velocity	252	12.3968	2.9849	6.0000	19.0000
Alter	252	16.0952	3.1202	3.0000	20.0000
Sediment	252	11.5119	3.9835	2.0000	19.0000
Riffl_Sinu	252	15.3611	3.8415	3.0000	20.0000
Chanflow	252	14.9603	3.0699	6.0000	20.0000
BnkStbTot	252	13.9960	3.9524	2.0000	20.0000
BnkVegTot	252	13.5079	4.1221	1.0000	20.0000
RipVegTot	252	12.3095	5.7268	0	20.0000
Total	252	136.8492	23.3214	78.0000	191.0000
Temp	248	17.0648	3.8846	6.9800	28.1000
pH	245	7.2178	0.8714	3.0799	10.0699
DO	248	9.1532	1.4463	5.5700	19.3999
Conduct	248	207.4032	222.0963	13.0000	1530.000
Fecal	246	836.7032	1928	0	12000.00
Acid_Hot	199	5.7190	5.5073	1.0000	67.0000
Alkalinity	199	50.4357	61.6067	5.0000	560.0000
Hardness	183	78.8642	73.9390	2.6199	436.2000
Sulfate	198	47.5236	74.7286	5.0000	380.0000
Chloride	183	6.9257	17.5157	1.0000	222.0000
Tot_Phos	213	0.0387	0.1146	0.0100	1.2900
NO2_NO3_N	213	0.3733	0.8207	0.0200	10.8000
Al_Tot	199	0.4033	0.8073	0.0200	8.1000
Ca_Tot	184	20.7196	19.3843	0.4400	122.0000
Fe_Tot	199	0.4013	0.4016	0.0300	2.5999
Mg_Tot	186	6.3894	6.7460	0.3700	36.7999
Mn_Tot	195	0.0869	0.1639	0.0030	1.2200

6.6.2 Hotspot for the WV 2004 data

After data manipulation, we used the multivariate hotspot detection modeling approach (MHDM) to find the hotspot which has the strongest multivariate linear relationship between responses and potential stressors. We ran 10,000 Voronoi tessellations for the 2- to 5-cluster partitions to find the hotspot and the corresponding model structure. The minimum cluster size was constrained to be 38, which is 15% of the total observations. With the model selection procedure incorporated inside the partition procedure, the multivariate hotspot detection modeling approach gave us the maximum R-square-like values and the model structures within the hotspots for 2- to 5-cluster partitions. Figure 6.7 displays the locations of hotspots detected for different sizes of partitions. The red triangles are the observations in the hotspot region. Hotspot regions detected from 3- to 5-cluster partitions are approximately located in the same area of northeast West Virginia. Table 6.11 summarizes the model and R-square-like quantity for different sizes of partitions. Cluster 2 in both the 4-cluster and 5-cluster partition solution gives us the same cluster and model. The region of this cluster is the global hotspot and the biological responses have the strongest linear relationship with Conduct, pH, Mn (heavy metal), cover (plant coverage), sedimentation and Hardness (dissolved substances).

Figure 6.7 Hotspots detected (red area) for 2- to 5-cluster partitions for the dataset sampled from West Virginia in 2004.



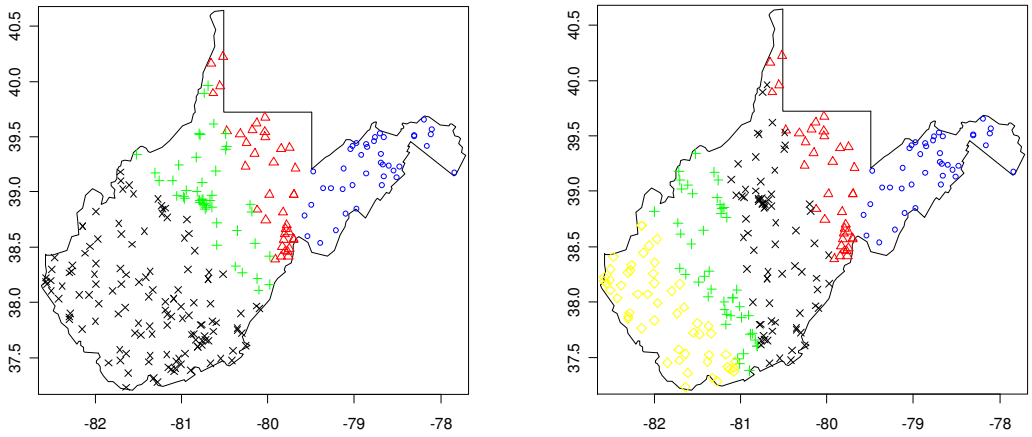


Table 6.11 Hotspot detection results using a MHDM approach for the data sampled in West Virginia in 2004. The minimum number of observations per cell is 38.

Number of partitions	R-Square-like value/obs	Maximum R-quare-like	Model within
1	**0.2838/252**	0.2838	y ~ Sulfate + Embed + Chloride + pH + Alkalinity + Mn_Total
2	**0.6437/38**	0.6437	y ~ Sulfate + Embed + pH + Alkalinity + Chloride + Riffl_Sinu
	0.2984/214		
3	**0.6822/39**	0.6822	y ~ Conduct + Ca_Tot + Cover + M_Total + Sediment + Al_Tot
	0.2588/165		
	0.4323/48		
4	0.5029/40	**0.7165**	**y ~ Conduct + pH + Sediment + Cover + Hardness + Mn_Total**
	0.7165/38		
	0.5096/46		
	0.2624/128		
5	0.5029/40	**0.7165**	**y ~ Conduct + pH + Sediment + Cover + Hardness + Mn_Total**
	0.7165/38		
	0.4875/46		
	0.3315/53		
	0.2534/53		

When we look at the model within the global hotspot, we find that conductivity and hardness are highly correlated to each other and therefore have high inflation values (inflation factor>10) in the RDA analysis. One of them was dropped to get a better model. Table 6.12 gives the inflation values for the model in the hotspot before and after we dropped Hardness.

Table 6.12 The inflation factor values for the model in the hotspot of the WV 2004 data.

Varible	inf factor before	inf factor after
Cover	1.3009	1.2862
Sediment	1.7393	1.5206
pH	2.5168	1.5591
Conduct	**16.0451**	2.0308
Hardness	**21.0167**	--
Mn_Total	2.4467	2.0053

6.6.3 RDA result(s) and discussion within hotspot

The result of redundancy analysis using the model selected by dropping the hardness for the WV 2004 data sampled within the hotspot region is illustrated in Table 6.13. Figure 6.8 is the triplot for this RDA analysis. The sum of the canonical eigenvalues for the analysis is 0.682. The total variance for the biotic community is 1. Thus the R-square-like quantity for this regional model is 0.682. The percentage of the total prediction variance explained by the first two axes is 96.4%. The 999 permutation tests indicate the significance of the model relationship within the hotspot region. The first axis is positively associated with the heavy metal Mn, Conductivity and negatively associated with Cover and Sediment. These environmental variables are correlated with the first RDA axis and describe a gradient from sites with greater sediments and cover at the negative end to sites with greater heavy metal concentration and conductivity in the positive end. This axis is associated with increasing mining effects. The second axis is closely correlated with pH.

Figure 6.9 gives the details of the estimated linear relationship between the biotic metrics and the scores for the first axis. EPT_Taxa and Total_Taxa have negative linear

relationships with this axis and HBI is positively related with it. In other words, we can say that high mining effects will increase HBI and decrease EPT_Taxa and Total_Taxa. The percentage of the 2 dominant Taxa and the percentage of Chrionomidae have positive relationships with the first axis and the percentage of EPT is negatively related with it. Thus, we can say that high mining effects increase the percentage of the 2 dominant Taxa and percentage of Chrionomidae and decrease the percentage of EPT Taxa. All of these findings in the analysis are consistent with the findings of the previous hotspot analysis using the West Virginia data from 1996 to 2005.

Table 6.13 Results of the redundancy analysis for the hotspot region of the WV 2004 data.

Statistics	Axis1	Axis2	Axis3	Axis4	Total variance
Eigenvalues	0.487	0.171	0.021	0.002	1.000
Species-environment correlations	0.865	0.828	0.666	0.358	
Cumulative percentage variance					
of species data	48.7	65.8	67.9	68.1	
of species-environment relation	71.4	96.4	99.5	99.9	
Sum of all eigenvalues					1.000
Sum of all canonicaleigenvalues					0.682
Test of significance of all canonical axes : P-value* =					0.0020
Test of significance of first canonical axis: P-value* =					0.0020

* The P value is the result of 999 permutation tests

Table 6.14 gives the details on how each biological metric (response) changes linearly with each stressor as illustrated in the two dimensional triplot in Figure 6.9. It also gives the R-square values for the individual response after the RDA analysis. All the R-square values have significantly increased within the hotspot region.

Table 6.14 Estimated linear relationship between the species score and the environmental score for the first two axes and the R-square value for individual responses (n=38).

Regression coefficients: Transformed Y's are regressed on Transformed X's						
	P2Dom	P_Chiro	P_EPT	HBI	EPT_Taxa	Total_Taxa
Cover	-0.2348	-0.2505	0.1853	-0.2422	0.2399	0.1709
Sediment	-0.0200	0.3782	-0.3104	0.4017	0.0974	0.0991
pH	-0.7334	-0.2306	-0.0905	0.3615	0.1356	0.4846
Conduct	0.4702	0.7479	-0.6936	0.4129	-0.5118	-0.6916
Mn_Total	-0.0341	0.1055	-0.1763	0.1392	-0.3339	-0.0571
R squared:						
	0.6335	**0.6423**	**0.7204**	**0.6039**	**0.7833**	**0.7088**

Figure 6.8 Triplot of the hotspot region for the WV 2004 data.

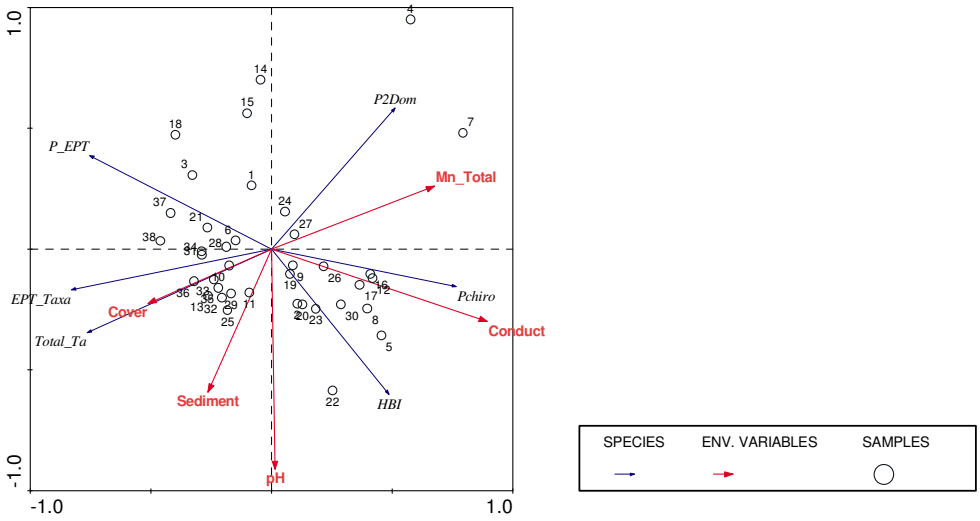
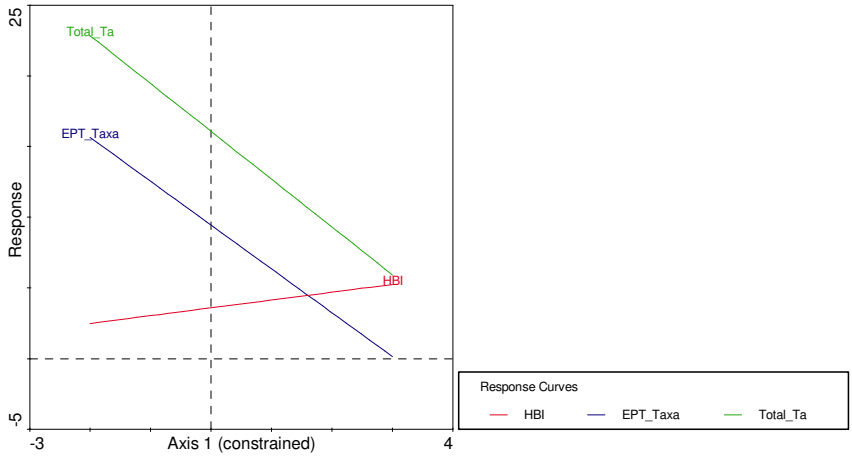
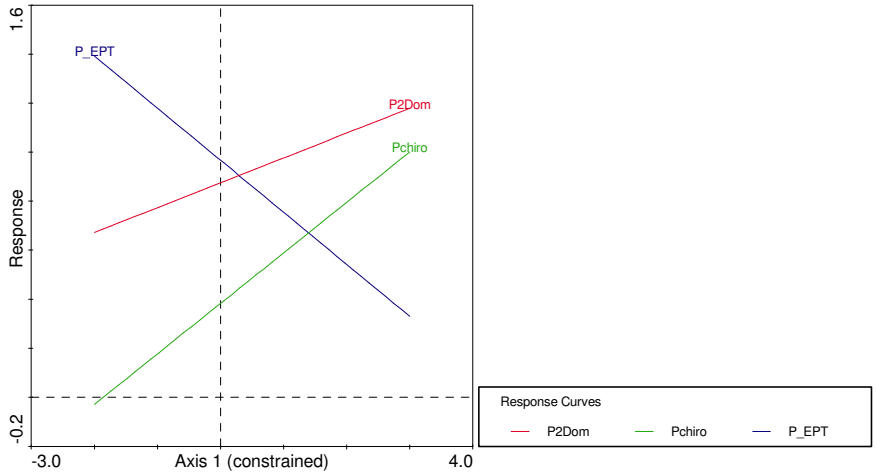


Figure 6.9 Estimated linear relationships between the biotic metrics and the first axis within the hotspot region in the WV 2004 data.





6.7 Summary and discussion

6.7.1 Coal Mining Effects on biological species

Coal accounts for one-third of the total energy used and over one-half of the electricity generated in the United States. Coal from the Appalachian Basin fields in West Virginia is a major resource for the Nation (USGS, 2006). During the past century, coal mining activities in West Virginia have contaminated streams by increasing acidity and metal concentration in sediments. Acid mine drainage (AMD) from coal mining has been identified as the factor having the most widespread effect on water quality in the Allegheny and Monongahela River Basins. Most of the drainage is from abandoned mine sites. Water discharging from deep-mine openings and surface mine seeps results in increased concentrations of acidity, iron, manganese, aluminum, and sulfate in receiving streams and rivers (Sams III & Beer, 2000). A study by Williams and others (1996) on 270 mine discharges in the Stonycreek River Basin found that many of the discharges had a pH less than 3.0. Water samples from these discharges generally had high concentrations of iron, manganese, aluminum, sulfate and high acidity. Water quality is thus severely degraded when mine discharges enter streams and rivers. AMD can seriously affect aquatic habitats resulting in stream bottoms covered with orange or yellow brown iron oxide or white aluminum oxide precipitates. The combined effect of chemical and physical stressors on watershed ecosystems is a decline in ecosystem health, thus the loss of biodiversity (fish, macroinvertebrates, algae). AMD has eliminated fish completely from some rivers and streams, and others support only a few acid tolerant benthic species such as the family of Chironomidea. In our study, the AMD effect will result in a decrease in Total Taxa, EPT_Taxa (Ephemeroptera Plecoprtera Trichoptera Taxa) and P_EPT and an increase in the metrics %Chironomidea, P_2Dom and HBI .

6.7.2 Summary of findings using MHDM

Figure A6.2 in the appendix is the area of coal mining in the Allegheny-Monongahela River Basins studied by Sams III and Beer (2000). According to this study, the watersheds in the Allegheny-Monongahela basin within West Virginia were more

seriously affected by the AMD and had twice the median sulfate concentration than did the area in Pennsylvania. Our RMPM hotspot detection approach helped us locate this Allegheny-Monongahela basin area in West Virginia. The hotspots we found using all observations from 1996 to 2005 as well as those from 2004 only belong to the West Virginia part of the Allegheny-Monongahela River Basin.

For the analysis using the 1995-2005 observations, the hotspots detected for the 3, 4 and 5 cluster partitions are approximately located in the same spot of the northeast side of West Virginia (Figure 6.2). Although the models within the hotspots are slightly different (Table 6.3), they all include conductivity, metal, PH and sediments which are indicators of the AMD effect. The hotspot in the 5 cluster partition, which is the optimal global hotspot detected, further included anion, the combination of Sulfate and Chloride, an important indicator of AMD effect. For the analysis of the year 2004 samples only, the hotspot detected for the 3, 4 and 5 cluster partitions are approximately located in the same spot of the northeast side of West Virginia (Figure 6.7). The models within the hotspots for the 4 and 5 partition solutions are the same (Table 6.11). The above results indicates that the MHDM approach is able to detect the hotspot quickly from a smaller cluster size partition, say 3 in this case, as long as the minimum number of observations within a sub-cluster is adequate.

The hotspot region detected using the year 2004 observations is consistent with the region detected using the 1995-2005 observations except that the delineation of hotspot for 2004 is closer to that of the Allegheny-Monongahela River Basin in West Virginia. The possible reason for this is that the missing values for the observations in 2004 are less and we used all of the original variables in our model while we used merged water chemistry variables for the 1996-2005 analysis.

Both RDAs of the hotspots for 2004 and 1996-2005 observations have rather similar results. The first ordination axis is positively associated with heavy metals, conductivity and negatively associated with vegetation coverage and sedimentation. It represents the mining (AMD) effect within the detected hotspot. Biodiversity variables

such as EPT_Taxa, percentage of EPT and Total_Taxa have an estimated negative linear relationship with this axis. The acid tolerant species and indices such as the percentage of the 2 dominant Taxa, percentage of Chrionomidae and HBI are positively related to the axis. In the hotspot area, we found that a high mining effect decreases the biodiversity and increases the acid tolerant metrics.

Appendix

Figure A6.1 The area of coal mining in the Allegheny-Monongahela River Basin as studied by Sams III and Beer in 2000 (Pictures from Sams III & Beer, 2000).

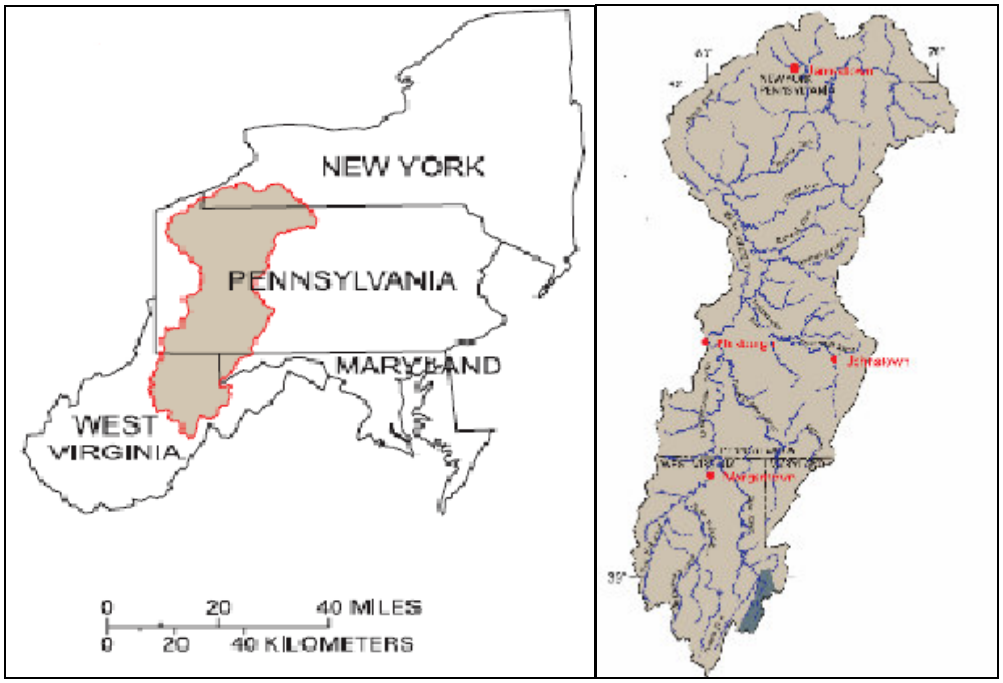


Table A6.1 The summary of the WV 2004 data.

Variable	N	Mean	Std Dev	Minimum	Maximum
P_2Dom	252	58.3146	15.6782	29.72	98.0000
P_Chiro	252	21.8810	21.1124	0.4200	94.3300
P_EPT	252	62.8031	23.2743	0.9100	96.2000
HBI	252	3.9479	0.9785	0.7200	6.0400
EPT_Taxa	252	9.1547	3.4795	2.0000	18.0000
Total_Taxa	252	15.8055	4.2103	3.0000	27.0000
WVSCI	252	72.8382	17.1977	19.5000	97.4300
Cover	252	13.8055	3.1770	2.0000	19.0000
Embed	252	12.9206	3.7108	2.0000	20.0000
Velocity	252	12.3968	2.9849	6.0000	19.0000
Alter	252	16.0952	3.1202	3.0000	20.0000
Sediment	252	11.5119	3.9835	2.0000	19.0000
Riffl_Sinu	252	15.3611	3.8415	3.0000	20.0000
Chanflow	252	14.9603	3.0699	6.0000	20.0000
BnkStbTot	252	13.9960	3.9524	2.0000	20.0000
BnkVegTot	252	13.5079	4.1221	1.0000	20.0000
RipVegTot	252	12.3095	5.7268	0	20.0000
Total	252	136.8492	23.3214	78.0000	191.0000
Temp	248	17.0648	3.8846	6.9800	28.1000
pH	245	7.2178	0.8714	3.0799	10.0699
DO	248	9.1532	1.4463	5.5700	19.3999
Conduct	248	207.4032	222.0963	13.0000	1530.000
Fecal	246	836.7032	1928	0	12000.00
Acid_Hot	199	5.7190	5.5073	1.0000	67.0000
Alkalinity	199	50.4357	61.6067	5.0000	560.0000
Hardness	183	78.8642	73.9390	2.6199	436.2000
Sulfate	198	47.5236	74.7286	5.0000	380.0000
Chloride	183	6.9257	17.5157	1.0000	222.0000
Tot_Phos	213	0.0387	0.1146	0.0100	1.2900
NO2_NO3_N	213	0.3733	0.8207	0.0200	10.8000
Al_Tot	199	0.4033	0.8073	0.0200	8.1000
Ca_Tot	184	20.7196	19.3843	0.4400	122.0000
Fe_Tot	199	0.4013	0.4016	0.0300	2.5999
Mg_Tot	186	6.3894	6.7460	0.3700	36.7999
Mn_Tot	195	0.0869	0.1639	0.0030	1.2200

Table A6.2 Correlation metrics of the WV 2004 data.

Correlation of SCI metrics

	P2Dom	P_Chiro	P_EPT	HBI	EPT_Taxa	Total_Taxa
P2Dom	1.0000	0.5980	-0.4534	0.4552	-0.6192	-0.6801
P_Chiro	0.5980	1.0000	**-0.8277**	0.6778	-0.4376	-0.4644
P_EPT	-0.4534	-0.8277	1.0000	-0.7328	0.4914	0.3759
HBI	0.4552	0.6778	-0.7328	1.0000	-0.5781	-0.4605
EPT_Taxa	-0.6192	-0.4376	0.4914	-0.5781	1.0000	**0.8551**
Total_Taxa	-0.6801	-0.4644	0.3759	-0.4605	0.8551	1.0000
WVSCI	**-0.8115**	**-0.8303**	**0.8066**	**-0.7647**	**0.8141**	**0.7859**

Correlation of habitat variables

	Cover	Embed	Alter	Sediment	Riffl_ Sinu	Bnk VegTot
Cover	1.0000	**0.7102**	0.1767	**0.6144**	**0.5939**	0.4274
Embed	**0.7102**	1.0000	0.1376	**0.7625**	**0.6530**	0.2534
Alter	0.1767	0.1376	1.0000	0.1403	-0.0167	0.5242
Sediment	**0.6144**	0.7625	0.1403	1.0000	**0.5849**	0.3732
Riffl_Sinu	**0.5939**	**0.6530**	-0.0167	0.5849	1.0000	0.2219
BnkVegTot	0.4274	0.2534	**0.5242**	0.3732	0.2219	1.0000

Correlation of field variables

	Temp	pH	Conduct
Temp	1.0000	0.1752	0.1732
pH	0.1752	1.0000	**0.5401**
Conduct	0.1732	0.5401	1.0000

Correlation of lab variables

	Akalini	Hardness	Sulfate	Chloride	Al_Tot	Ca_Tot	Mg_Total	Mn_Total	Fe_Tot
Alkalini	1.0000	**0.8219**	**0.5223**	**0.5159**	-0.0966	**0.8378**	**0.7135**	-0.0376	0.0578
Hardness	**0.8219**	1.0000	**0.7971**	**0.6175**	0.0796	**0.9843**	**0.9388**	0.2665	0.1525
Sulfate	**0.5223**	**0.7971**	1.0000	0.6153	0.1865	**0.7109**	**0.9020**	0.4429	0.1710
Chloride	0.5159	0.6175	0.6153	1.0000	-.0498	0.**5952**	**0.6190**	0.2084	0.0551
Al_Tot	-.0966	0.0796	0.1865	-.0498	1.0000	0.0746	0.0847	**0.5242**	**0.7501**
Ca_Tot	**0.8378**	**0.9843**	**0.7109**	**0.5952**	0.0746	1.0000	**0.8700**	0.2274	0.1531
Mg_Total	**0.7135**	**0.9388**	**0.9020**	**0.6190**	0.0847	0.8700	1.0000	0.3267	0.1448
Mn_Total	-.0376	0.2665	0.4429	0.2084	**0.5242**	0.2274	0.3267	1.0000	0.5662
Fe_Tot	0.0578	0.1525	0.1710	0.0551	**0.7501**	0.1531	0.1448	0.5662	1.0000

7 Summary and Future Research

The spatial partition modeling approach and the adjusted BCART method provide two alternatives for partitioning a geographical/spatial region into disjoint clusters when the focus is on response-stressor relationships. Both methods have advantages over the benchmark model in terms of performance criteria such as prediction accuracy, likelihood, and R-square. When applied to a dataset, the adjusted BCART method provides a better partition than the usual BCART by placing the observations more evenly over the space of interest, and thus gives BCART more flexibility in application. The spatial partition modeling approach uses Voronoi diagrams and a Monte Carlo simulation technique to search for the optimal partition over the space of interest. This approach has the flexibility to be extended to many modeling situations; we explored multicategorical regression and multivariate regression in our research.

We need to point out that because of the very nature of the Bayesian approach, it is hard to deal with the high dimensional multivariate analysis using the adjusted BCART method. In such cases, the spatial partition approach is the only choice. In the univariate modeling situation, both methods can be used to partition the space and the partition criterion is chosen according to the research interest. Our research shows that the applications of both methods have improved performance compared to the benchmark models in terms of the performance criterion value of the interest. Although both methods are good, there are many opportunities for improvement. The following is several thoughts on the spatial partition modeling approach:

1. For the Voronoi diagram based clustering approach, the methods proposed are ad-hoc. The proposed model selection criteria need to be evaluated by using simulation if its properties can not be analytically evaluated.

2. For two sites near the boundary of clustering, which are close to each other but are in different clusters, the models for them could be very different. But since they are very close, we expect the predictions for those two sites to be close.

Smoothing splines or other adaptive smoothing techniques can be explored after the clustering.

3. In this approach, we assumed that the data within different clusters are independent. This assumption is too strict in the sense that the spatial correlation exists among different clusters in general. How best to fit the model within each cluster by including spatial dependence inside the modeling procedure is another area of work.

Bibliography

Agresti, A. (2002), "Categorical Data Analysis", John Wiley & Sons.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", Second International Symposium on Information Theory, pp. 267-281.

Anderson, C. J. (2006), "Applied Categorical Data Analysis", Course notes of EdPsych 590AT/Psych 593 CA, http://www.ed.uiuc.edu/courses/EdPsy490AT.

Banfield, J. D.and Raftery, A. E. (1993), "Model based Gaussian and Non-Gaussian Clustering", Biometrics 49:803-821.

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

Breiman, L., Friedman, J. H., Olshen, R.A. and Stone, C. J. (1984). "Classification and Regression Trees", Wadsworth & Brooks/Cole Advanced Books & Software.

Burnham K. P. and Anderson, D. R. (1998), "Model Selection and Inference: a Practical Information-theoretic Approach", Springer.

Chipman, H. A., George, E. I., McCulloch, R. E. (1998), "Bayesian CART Model Search", Journal of the American Statistical Association 93:936-948.

Chipman, H., George, E. I., and McCulloch, R. E. (2000), "Hierarchical Priors for Bayesian CART Shrinkage", Statistics and Computing 10(1):17-24.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2003), "Bayesian Treed Generalized Linear Models", Bayesian Statistics 7:85-104, Clarendon Press, Oxford.

Davies, P. T. and Tso, M. K. (1982), “Procedures for Reduced-rank Regression”. Applied Statistics 31:244-255.

De’ath, G. and Fabricius, K. E. (2000), “Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis”, Ecology 81(11):3178-3192.

Denison, D. G. T., Mallick, B. K. and Smith A. F. M. (1998), “A Bayesian CART Algorithm”, Biometrika 85(2):363-377.

Denison, D. G. T and Holmes, C. C. (2001), “Bayesian Partitioning for Estimating Disease Risk”, Biometrics 57(1):143 -149.

DeSarbo, W. S. and Cron, W. L. (1988), “A Maximum Likelihood Methodology for Clusterwise Linear Regression”, Journal of Classification 5(2):249–282.

Fawcett, T. (2003), “ROC Graphs: Notes and Practical Consideration for Data Mining Engineers", Technical Report HPL-2003-2004.

Fraley, C. and Raftery, A. E. (2002), “Model-based Clustering, Discriminant Analysis and Density Estimation”, Journal of the American Statistical Association 97(458):611–631.

Gao, F. (2005), “Credit Risk Analysis Using Bayesian Classification and Regression Tree”, Internship report, Virginia Tech.

Green, J., Passmore, M. (2000), “A Survey of the Condition of Streams in the Primary Region of Mountaintop Mining/Valley Fill Coal Mining”, US Environmental Protection Agency Region III report.

Griffith, M. B., Kaufmann, A. T. , Herlihy A. T., Hill B.H. (2001), “Analysis of Macro Invertebrate Assemblages in Relation to Environmental Gradients in Rocky Mountain Streams”, Ecological Application 11(2):489-505.

Godinez-Dominguez, E., Freire, J. (2003), “Information-theoretic Approach for Selection of Spatial and Temporal Models of Community Organization”, Marine Ecology Progress Series 253:17-24.

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

Hawkins, D. S., Allen, D. M. and Stromberg, A. J. (2001), “Determining the Number of Components in Mixtures of Linear Models”, Computational Statistics and Data Analysis 38(1):15-48.

Hoffmann, K. A. and Chiang, S. T. (1993), “Computational Fluid Dynamics for Engineering”, Engineering Education System.

Holmes, C. C., Denison, D. G. T. and Mallick, B. K. (2001), “Bayesian Partitioning for Classification and Regression”, Biometrics 57(1):143-149.

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.

Lamon, E. C. and Stow, C. A. (2004), “Bayesian Methods for Regional-Scale Eutrophication Models”, Water Research 38(11):2764-2774.

Lawal, B. (2003), “Categorical Data Analysis with SAS and SPSS Applications”, Lawrence Erlbaum Associates, London.

Lee, D. T., and Schacter, B. J. (1980), “Two Algorithms for Constructing a Delaunay Triangulation”, International Journal of Computer and Information Sciences 9(3):219 - 242.

Legendre, P. and Legendre, L. (1998), “Numerical Ecology”, 2nd English Edition. Elsevier Science BV, Amsterdam.

Leisch, F. (2004), "Exploring the Structure of Mixture Model Components", Compstat' 2004 Symposium:1405-1412. Phisika Verlag, Springer.

Lejeune-Dirichlet G. (1850), "Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen", Journal für die Reine und Angewandte Mathematik 40:209-227.

Lipkovich, I. (2003), "Bayesian Model Averaging and Variable Selection in Multivariate Ecological Models", Ph.D dissertation, Virginia Tech.

Magidson, J., Eagle, T and Vermunt, J. (2003), "New Developments in Latent Class Choice Models", Sawtooth Software Conference Proceedings:89-112.

McLachlan, G. J. and Peel, D. (2000), "Finite Mixture Models", Wiley, New York.

McCullagh M. J. (2006), "Detecting Hotspots in Time and Space", the 23rd meeting of the ISO/TC211 Plenary and Working Groups, Riyadh, SA

Morgan, J. and Sonquist, J. (1963), "Problems in the Analysis of Survey Data and a Proposal", Journal of the American Statistical Association 58(302):415-434.

Muhamma, R., http://www.personal.kent.edu/~rmuhamma/Compgeometry/MyCG/CG-Applets/VoroDiagram/vorocli.htm.

Oksanen J. (2006) "Multivariate Analysis of Ecological Communities in R: Vegan Tutorial", http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf.

Olsen, E. C. B., (2003) "Modeling Slow Lead Vehicle Lane Change", Ph.D dissertation, Department of Industrial Engineering, Virginia Tech.

Palmer, M., Webpage: http://ordination.okstate.edu/.

Patil G. P. and Taillie C. (2003), "Geographic and Network Surveillance via Scan Statistics for Critical Area Detection", Statistical Science 18(4):457–465.

Pepe, M.S., Cai, T. and Longton G. (2005), "Combining Predictors for Classification Using the Area Under the Receiver Operating Characteristics Curve", Biometrics 62 (1):221-229.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research (with discussion)", Sociological Methodology 25:111-163.

Renka, R. J. (1996), "Algorithm 751: TRIPACK: a Constrained Two-Dimensional {Delaunay} Triangulation Package", ACM Transactions on Mathematical Software 22: 1-8.

Rissanen, J. (1986), "Order Estimation by Accumulated Prediction Errors", Journal of Applied Probability 23:55-61.

Sams III, J. I. and Beer, K. M. (2000), "Effects of Coal-Mine Drainage on Stream Water Quality in Allegheny and Monongahela River Basins—Sulfate Transport and trends", USGS Water-Resources Investigations Report 99-4208.

Sclove, S. L. (1993), "Small- and Large-Sample Statistical Model Selection Criteria", Proceedings of Fourth International Workshop on Artificial Intelligence and Statistics.

Smith, E. P. (2003), "Model Based Clustering for Classification of Aquatic Systems and Diagnosis of Ecological Stress", EPA STAR Grant proposal.

ter Braak, C. J. F. (1986), **"**Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis**",** Ecology 67(5):1167-1179.

ter Braak, C. J. F. (1987), "Unimodal Models to Relate Species to Environment" , DLO-Agriculture Mathematics Group, Wageninggen.

ter Braak, C. J. F. and Smilauer, P. (1998), "CANOCO Reference Manual and User's Guide to CANOCO for Windows: Software for Canonical Community Ordination (version 4)", Microcomputer Power, Ithaca, New York, USA.

Thieling, T. M. (2006), "Assessment and Predictive Model for Brook Trout Population status in Eastern United States", Master thesis, Deptartment of Biology, James Madison University.

Tibshirani, R., Walther, G. and Hastie, T. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic", Journal of Royal Statistical Society 63(2):411-423.

Voronoi G. (1907). "Nouvelles applications des paramètres continus à la théorie des formes quadratiques", Journal für die Reine und Angewandte Mathematik 133:97-178.

Wedel, M. and DeSarbo, W.S. (1995), "A Mixture Likelihood Approach for Generalized Linear Models", Journal of Classification 12(1):21–55.

Williams, D. R., Sams III, J. I., and Mulkerrin, M. E. (1996), "Effects of Coal Mine Discharges on the Quality of the Stonycreek River and Its Tributaries, Somerset and Cambria Counties", USGS Water-Resources Investigations Report 96-4133.

Yuan, L. L. and Norton, S. B. (2004), "Comparing Responses of Macroinvertebrate Metrics to Increasing Stress", Journal of the North American Benthological Society 22 (2):308–322.

Zhang, H., Thieling, T, Bates Prins, S.C., Smith, E.P., Hudy, M. (2008), "Model-based Clustering in a Brook Trout Classification Study within the Eastern United States", Transactions of the American Fisheries Society 137:841-851.