A Finite Mixture Approach for Identification of Geographic Regions with Distinctive Ecological Stressor-Response Relationships

David Farrar[1,2]

Samantha C. Bates Prins[1]

Eric P. Smith[1]

[1]Department of Statistics, Virginia Tech, Blacksburg, VA 24061

[2]To whom correspondence should be addressed (email: dfarrar@vt.edu)

Summary. We study a model-based clustering procedure that aims to identify geographic regions with distinctive relationships among ecological and environmental variables. We use a finite mixture model with a distinct linear regression model for each mixture component, relating a measure of environmental quality to multiple regressors. Component-specific values of regression coefficients are allowed, for a common set of regressors. We implement Bayesian inference jointly for the true partition and component regression parameters. We assume a known, prior classification of measurement locations into "clustering units," where measurement locations belong to the same mixture component if they belong to the same clustering unit. A Metropolis algorithm, derived from a well-known Gibbs sampler, is used to sample the posterior distribution. Our approach to the label switching problem relies on constraints on cluster membership, selected based on statistics and graphical displays that do not depend upon cluster indexing. Our approach is applied to data representing streams and rivers in the state of Ohio, equating clustering units to river basins. The results appear to be interpretable given geographic features of possible ecological significance.

Acknowledgements: We thank Gene Yagow for assistance in geographic analysis. Funding for this research was provided by U.S. EPA-Science To Achieve Results (STAR) Grant RD 83136801-0. The research has not been subjected to any USEPA review and so no USEPA endorsement should be inferred.

Keywords: cluster analysis, model-based clustering, finite mixture model, Bayesian statistics, Markov chain Monte Carlo, ecoregions, geographic information systems, environmental statistics.

1. Introduction

Recognition that the ecological effect of a pollutant or other environmental stressor may depend on regional influences such as climate and soils has stimulated efforts to delineate geographic regions useful for purposes of environmental analysis and decision making. Uses of such regions include stratification of sampling studies, and identification of appropriate ranges of inputs for mechanistic models.

Standard systems of "ecoregions" have been developed by governmental agencies. Ecoregions used in the U.S. are based on combinations of soil types, landform, land-use, climate and biological communities. McMahon *et al*. (2001) provide an account that integrates perspectives from multiple U.S. agencies. Standardized systems of ecoregions are based primarily on expert, subjective integration of diverse information, with statistical procedures recognized as playing an important supporting role. The ecoregion systems may facilitate inter-disciplinary decisions, accounting for social and economic as well as environmental considerations. While the systems provide default regions for many analyses, it is recognized that they may not be ideal for analysis of some specific environmental problems.

As an alternative (or complement) to the standard systems of ecoregions, based on expert judgement, statistical classification methods may be used to identify regions relating quantitative measures of environmental quality to measures of environmental stress (Lamon and Stowe, 2004; Robertson and Saad, 2003). Results from such approaches may corroborate the usefulness of standard ecoregions, or may suggest alternative divisions relevant to specific measured variables. A statistical perspective is that we may have a better chance of success with a simple

type of model such as linear regression, if such a model is assumed to hold within restricted geographic regions rather than globally, in view of the possibility for complex nonlinearities and interactions on a more global scale.

Here, we study the use of a statistical model for predicting a measure of biological environmental quality, allowing regional variation in the coefficients of a multiple linear regression model for prediction of a measure of environmental quality. We think that some support for a model with region-specific regressions is provided by the suggestion that ecoregions should be relatively homogeneous with respect to ecological processes (McMahon *et al*., 2001). The delineation of regions, as well as the values of regression parameters, are treated as unknowns subject to statistical inference procedures. We use a finite mixture (FM) model. With the FM approach, generation of an observation is modeled as a two-stage process in which the first stage samples from K subpopulations (termed "components") and the second stage generates observable values by sampling component-specific models. In our situation, the component-specific models are linear regression models. We use flexible Bayesian inference procedures that provide a probabilistic treatment of uncertainty in the delineation of regions, as well as for parameters of the regression model. We use somewhat conventional priors and sample the posterior using a Markov chain Monte Carlo (MCMC) procedure (see also Viele and Tong (2002), and sources indicated therein).

If a method of cluster analysis method is to be useful in identification of ecoregions, a desirable feature is a preference for placing in the same cluster locations that are near together in space. Our current approach for implementing a preference of this type was proposed in Lipkovich *et al*. (to appear), who relied on a type of classification model rather than a finite mixture model. According to this approach we rely upon fixed groups of nearby measurement locations, which we term clustering units (CUs). Measurement locations belonging to the same

CU are constrained to belong to the same cluster. In the FM model framework the components may be conceived as populations of CUs rather than populations of measurement locations.

We present a 2-component model based on a water quality data set for the state of Ohio. For that application the CUs are equated to river basins and the measurement units to stream segments within basins. Although our method does not force spatial contiguity of the clusters, we find some correspondence between our clusters and important geographic features of the state.

In our application, we found rapid mixing of a posterior sampling algorithm, and no evidence of local optima. These results seem to contrast with pessimistic statements in the literature of FM models related to the tendency of MCMC algorithms to become trapped in neighborhoods of local optima (Celeux *et al*., 2000; Viele and Tong, 2002; Ter Braak *et al*., 2003). We suggest that the apparent lack of complications in our application is due to imposition of a number of constraints on model unknowns.

In model-based clustering a model identification issue arises from the fact that the likelihood is unchanged by permutation of the class indices (Richardson and Green, 1997; Celeux *et al*., 2000; Stephens, 2000; Lipkovich *et al*., to appear). In an MCMC implementation the same partition may occur multiple times with the cluster indices permuted, within the same chain or from one simulation to the next. This problem is often termed the "label switching" problem in the FM model literature. If label switching is not taken into account, some posterior summaries may be meaningless. Some MCMC convergence diagnostics rely on comparison of independent chains, and require that the components be indexed in a way that provides valid comparisons. For purposes of our application we rely on an identification constraint that fixes the assignment of K CUs to K distinct components. Also, we suggest uses of estimated probabilities that pairs of CUs belong to the same component. These probability estimates do not depend on the indexing of components.

Previous work on finite mixtures of regressions includes maximum likelihood based on the Estimation-Maximization (EM) algorithm (DeSarbo and Cron, 1988; Wedel and Kamakura, 2000). Viele and Tong (2002) provide a recent review and update of the Bayesian approach with MCMC implementation. Related approaches include modeling with finite mixtures of multivariate normal or multivariate *t* distributions. Some authors emphasize maximum likelihood by the EM algorithm (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 1998, 2002; McLachlan and Peel, 2000). Bensmail *et al*. (1997) explore Bayesian inference for mixtures of multivariate normals with Gibbs implementation. An ecological application involves classifying lake communities based on a mixture of multivariate normal distributions, implemented with a Gibbs sampler (Ter Braak *et al*., 2002). Our approach differs in that we treat as fixed, rather than random, the values for a subset of the variables (our regressors). Our approach has much in common with recent Bayesian treed model methodology, allowing leaf-specific regressions and relying on Metropolis implementations (Chipman, George, and McCulloch, 1998, 2002; Denison, Mallick, and Smith, 1998; Lamon and Stowe, 2002).

We initially describe our finite mixture model, with only minimal reference to inferential issues, as a model for the generation of environmental data. We then describe our Bayesian inference approach including our priors and our posterior sampling procedure. We present an application involving prediction of the Index of Biotic Integrity for Ohio Streams.

2. Finite Mixture Model with Component Linear Regressions

For each variable a single value is recorded, at each of n measurement locations. The measurement locations are grouped into B categories termed clustering units (CUs), denoted $CU_1,...,CU_B$. We assume measurements for n_i measurement locations in CU_i. The CUs are units for clustering in the sense that measurement locations belonging to the same CU are constrained to belong to the same cluster. Thus the output of the clustering procedure is a partition of the set of CUs into K categories. (For our application to Ohio water quality data river basins are

equated to CUs.) We treat K as known and consider models with different values of K for given data.

For purposes of a model-based clustering approach we assume that the data are drawn from K true classes which we term "components." (We reserve the term "cluster" for a category defined by a specific partition of observations, for example the set of observations assigned to the same component.) In a model-based approach, the component to which a CU belongs is an unknown subject to statistical inference. A variant of model-based clustering can be based on a finite mixture (FM) model. We may represent the true, unknown component for CU_i using a vector of unobserved indicators $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$ where z_{ik} equals 1 or 0 according as CU_i does or does not belong to the kth component. In the FM approach the indicator vectors for the CUs, say $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, are realizations of independent multinomial random variables $Z_1, \ldots, Z_B$, identically distributed with parameters 1 and $(\tau_1, \ldots, \tau_K)$, where $\tau_i = E(Z_i)$. The parameters $\tau_1, \ldots, \tau_K$ are termed "mixing proportions."

We record values of the response variable and values of p regressors, at n_i measurement locations in CU_i. We assume a component-specific linear regression model with regression coefficient vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. Allowing for component intercepts, each is a column vector of length $p + 1$. Values of the response variable for CU_i are arrayed in a column vector $\boldsymbol{y}_i = (y_1, \ldots, y_{n_i})^{\mathrm{T}}$ and corresponding values of regressors in a matrix $\boldsymbol{X}_i$. We include a column of 1's in $\boldsymbol{X}_i$ to allow for the component intercept, so that the matrix is $n_i \times (p + 1)$. We use the same set of regressors for each component but each regression coefficient including the intercept is free to differ among components. Fixing the partition, the regression model applied to CU_i is $\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_{\gamma_i} + \boldsymbol{e}_i$ where $\gamma_i \in \{1, \ldots, K\}$ is the index of the component that includes CU_i. The error vector $\boldsymbol{e}_i$ is a realization of a random variable $E_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Here σ^2 is a common

residual variance. $\boldsymbol{0}$ and $\boldsymbol{I}$ are respectively a column vector of zeros and an identity matrix, of appropriate dimensions. For a pair of CUs, say the *i*th and *j*th, the corresponding error vectors E_i and E_j are independent.

Our unknowns subject to statistical inference thus include mixing proportions $\tau_1, \ldots, \tau_K$, component regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, a common residual variance σ^2, and a partition of the set of CUs.

With regression parameters (coefficients and residual variance) and mixing proportions fixed, the posterior probability that CU_i belongs to the *k*th cluster based on observing $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_B$ is

$$\hat{z}_{ik} = \frac{\tau_k \phi\left(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \sigma^2\boldsymbol{I}\right)}{\sum_{l=1}^{K} \tau_l \phi\left(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_l, \sigma^2\boldsymbol{I}\right)} \qquad (1)$$

where $\phi\left(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{V}\right)$ denotes the density of a normal distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V}$. We will refer to $\hat{z}_{ik}$ as a "membership probability."

To cluster the CUs it is natural to use estimated membership probabilities and assign a CU to the component that includes it, with highest estimated probability. In a maximum-likelihood approach, maximum likelihood estimators for the mixing proportions and regression parameters may be plugged into Expression (1) and an observation may be assigned to the component with the highest "plug-in" posterior probability. In a Bayesian approach, posterior component membership probabilities can be computed easily using posterior sampling, taking into account uncertainty in the values of parameters that might be fixed with a "plug-in" approach. However, Expression (1) has a role in a Gibbs sampler implementing the Bayesian approach. In that context the quantity is a function of the parameters for fixed data which we may write $\hat{z}_{ik}\left(\{\tau_k\}, \{\beta_k\}, \sigma^2\right)$.

3 Bayesian Approach with Posterior Sampling

The components of our Bayesian approach are a joint prior distribution for the unknowns and an algorithm for sampling the posterior. For the logarithm of σ^2 and the $K\ (p\ +\ 1)$ regression coefficients we assume a joint prior approximately constant on $\mathbf{R}^{1+K(p+1)}$, independent of the mixing proportions, apart from implicit modifications of the prior in the form of constraints that we may impose on the Metropolis implementation. (Constraints on the prior for our current implementations include sign constraints for regression coefficients and a minimum value that applies to each mixing proportion.) We require a minimum number of units per cluster adequate to result in a proper posterior.

A conventional prior for the mixing proportions is a Dirichlet distribution with the simplex in $K - 1$ dimensions as the support and parameters $(\alpha_1,...,\alpha_K)$. In our applications we have assumed $\alpha_1 = \cdots = \alpha_K\ =\ \alpha$, so that the mixing proportions are identically distributed, each with mean $1 / K$.

Before describing our current algorithm, it seems helpful to present a "straightforward Gibbs sampler." The algorithm alternates between random partition of the CUs into clusters, and random generation of regression parameters, conditioning on the current partition. Gibbs algorithms are well established for implementation of finite mixture models, as recently reviewed by Viele and Tong (2002). Unknowns are drawn in turn from distributions with other unknowns fixed, which is to say by sampling of "full conditional" distributions. In our situation the full conditional distributions are standard results. Our current algorithm was derived by modifying this Gibbs sampler, after encountering some difficulties in applying the latter.

We display below expressions for a single iteration of the straightforward Gibbs sampler, updating each unknown. Regarding notation for describing iterative computations, it seems helpful to use primes to indicate variables that change in value from one iteration to the next. In

order to distinguish successive updates, a single prime indicates the value carried from the previous iteration, while double primes indicate the value at the end of the iteration. For example $\boldsymbol{\beta}_k'$ is the value of $\boldsymbol{\beta}_k$ before the expressions are executed and $\boldsymbol{\beta}_k''$ the value afterwards. The transpose of a matrix $\boldsymbol{M}$ will be denoted $\boldsymbol{M}^{\mathrm{T}}$.

With these conventions the updating expressions for our initial Gibbs sampler are

$$\begin{aligned}
\gamma_i'' &\leftarrow \text{random} \sim \texttt{categorical}\left(\hat{z}_{i1}', \ldots, \hat{z}_{iK}'\right), \\
&\quad \text{for each } i \text{ independently;} \\
\left(\tau_1'', \ldots, \tau_K''\right) &\leftarrow \text{random} \sim \texttt{Dirichlet}\left(\alpha + n_1'', \ldots, \alpha + n_K''\right); \\
\sigma''^2 &\leftarrow \sum_{i=1}^{B} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_{\gamma_i'}'\right)^{\mathrm{T}} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_{\gamma_i'}'\right) / \left(\text{random} \sim \chi_n^2\right); \\
\boldsymbol{\beta}_k'' &\leftarrow \text{random} \sim N\left(\left(\boldsymbol{X}_k''^{\mathrm{T}} \boldsymbol{X}_k''\right)^{-1} \boldsymbol{X}_k''^{\mathrm{T}} \boldsymbol{y}_k'', \sigma''^2 \left(\boldsymbol{X}_k''^{\mathrm{T}} \boldsymbol{X}_k''\right)^{-1}\right), \\
&\quad \text{for each } k \text{ independently;} \\
\hat{z}_{ik}'' &\leftarrow \hat{z}_{ik}\left(\{\tau_k''\}, \{\boldsymbol{\beta}_k''\}, \sigma^2\right), \text{ for each } (i, k).
\end{aligned}$$

In words, we generate a partition by assigning each CU to a component at random, where $\hat{z}_{ik}$ is the probability of assigning CU_i to component k. We use the resulting counts of CUs per cluster as observed counts for updating the full conditional for the mixing proportions. Then, fixing the partition, we update the regression parameters by sampling the posterior of a standard Bayes linear model. Our expression for updating $\boldsymbol{\beta}_k$ uses a design matrix $\boldsymbol{X}_k$ and dependent column vector $\boldsymbol{y}_k$ formed by stacking, in the same order, design matrices and dependent vectors for the CUs in the current cluster k. Finally, we update the membership probabilities using Expression (1). We initialize by setting each of the BK membership probabilities $\{\hat{z}_{ik}\}$ equal to $1 / K$.

We allow constraints on the signs of regression coefficients, the ranges of mixing proportions, and the counts of locations per cluster. We assume a minimum count of locations for each cluster (combining the count of locations for CUs belonging to a cluster). Propriety of the posterior may dictate a minimum count in case of an improper prior for regression parameters (Roeder and Wasserman, 1997; Diebolt and Robert, 1994). Based on essentially an

argument in Roeder and Wasserman (1997), in order for the posterior to be proper for the regression parameters, the conditional posterior fixing the partition must be proper for each partition that we allow. Thus with a conventional improper prior we must have at least $p + 1$ measurement locations in each cluster. At least one cluster must include $p + 2$ or more locations. We require a larger count than the minimum for propriety of the posterior, based on a concern that design matrices $\boldsymbol{X}_1,...,\ \boldsymbol{X}_K$ may be ill-conditioned if the cluster sizes are too small. To implement constraints in the MCMC algorithm we simply repeat the series of updating expressions, on each iteration, until all constraints met.

For an iteration of the Gibbs algorithm described above, cluster assignments are updated for all CUs, followed by updating of other unknowns. Instability may result from allowing large changes of the partition from one iteration to the next, considering that most CUs may be re-assigned among components in the first step (see Viele and Tong, 2002). We suggest that this instability may be exacerbated in the event of multicollinearity or points of high influence. In our initial application the algorithm worked well for a 2-component model but not for 3- and 4-component models. One apparent difficulty was that the constraints were not met in a reasonable number of attempts because partitions that met the minimum cluster size were improbable, given values of other unknowns. Because of these difficulties a new algorithm was devised which worked well for 3- and 4-component models for our application, as well as for the 2-component model. (We checked the revised algorithm by comparing results to the initial algorithm for the 2-component model.)

Our current algorithm differs from the initial Gibbs sampler by two modifications. First, instead of updating the cluster assignments for all CUs at once, the algorithm is modified so that updating the cluster assignment of a single CU alternates with updating of the other unknowns (regression parameters, mixing proportions). Such a feature can be pseudo-coded as

for $CU_1,...,CU_B$ {
update the cluster assignment of CU_i;
update mixing proportions and regression parameters;
update membership probabilities
}

In contrast to our initial Gibbs sampler the cluster assignments for individual CUs are updated at a lower frequency than regression parameters and mixing proportions.

The CUs may be considered in systematic or random order. In our approach CUs are considered in an order that is random within cycles of length *B*, such that the cluster assignment of each CU is updated once per cycle. At the start of the simulation we generate a random permutation of the indices 1,...,*B*, which gives the order that CUs are considered for the first *B* iterations. On the (*B*+1)th iteration we generate a second random permutation giving the order for iterations *B*+1,...,2*B*, and so on.

Our second modification relates to the procedure for updating the cluster assignment of a single CU. One may sample from the full conditional as in the initial Gibbs sampler. However, in our current algorithm we use an explicit Metropolis step similar to that used by Lipkovich *et al*. (to appear). The Metropolis proposal is to move a randomly selected CU from its current cluster to another cluster. In the case of $K > 2$, the new cluster is chosen with equal probability from the $K - 1$ candidates. For a proposal to move CU_i from cluster *k* to cluster *l*, the acceptance probability based on multinomial full conditionals equals $\min \{ r, 1 \}$ where, based on Expression (1),

$$r = \frac{\hat{z}_{ik}}{\hat{z}_{il}} = \frac{\tau_k \phi \left(\boldsymbol{y}_i; \boldsymbol{X}_i \boldsymbol{\beta}_k, \sigma^2 \boldsymbol{I} \right)}{\tau_l \phi \left(\boldsymbol{y}_i; \boldsymbol{X}_i \boldsymbol{\beta}_l, \sigma^2 \boldsymbol{I} \right)}. \qquad (2)$$

With this modification it was relatively easy to satisfy the minimum cluster size.

It is valuable to observe from Expression (2) that as α is increased in value the mixing proportions effectively cancel in the ratio *r* so that partitions are compared based on likelihoods.

In effect, the result is a "classification likelihood" method (see the discussion). Updating of other unknowns does not depend on the mixing proportions.

To estimate the probability that CU_i belongs to cluster k from the posterior sample, we compute the fraction of the sample with CU_i belonging to cluster k. We can use the term Estimated Component Membership Probabilities (ECMPs) for these estimates and denote them $\{\tilde{z}_{ik}\}$. Using the ECMPs, assignment of CUs to clusters is naturally accomplished by assigning each CU to the component that includes it with highest posterior probability. To express uncertainty in classifying the ith CU we may use $1 - \max_k \{\tilde{z}_{i1}, \ldots, \tilde{z}_{iK}\}$, as suggested by Lipkovich *et al.* (to appear)**.**

The ECMPs depend on the indexing of clusters and are therefore susceptible to the label-switching problem. We suggest that useful information on the pattern of clustering can be obtained from the posterior sample based on quantities that do not depend on the cluster indexing, which we term Estimated Co-Clustering Probabilities (ECCPs). The ECCPs can be represented in a symmetric matrix giving in the ith row and jth column the fraction of the posterior sample for which CU_i and CU_j are assigned to the same cluster. An ECCP estimates the probabilities that a pair of CUs belongs to the same component. A graphical display that we find useful is a dendrogram such as our Figure 1, which uses the matrix of ECCPs as a similarity matrix. Also, as an approach to resolving the label-switching problem in some cases, in the next section we use ECCPs to select an identifying constraint on cluster composition.

4. Application to Analysis of the Index of Biotic Integrity for Ohio Streams

We illustrate our methods using a data set representing 303 measurement locations on streams in Ohio. The data analyzed here are a subset of those assembled by Dyer *et al*. (1998, DWCSW). Our response variable is the Index of Biotic Integrity (IBI), an index of the health of fish communities. From the data available we used the measurement locations with complete

data for IBI and four regressors. Our data represent 16 of the 23 basins identified by DWCSW. Our regressors (using variable names from DWCSW) are:

DO_MED = site median dissolved oxygen concentration (mg/L);
PH_MED = site median pH;
ZNTOT_MED = site median total zinc (μg/L); and
QHEI = Qualitative Habitat Evaluation Index.

We will usually refer to the regressors simply as DO, pH, zinc, and QHEI. IBI and QHEI are "multimetric indices" computed by summing together contributing "metrics." Each contributing metric has a range of 1-5 with a value of 5 representing the most favorable value (references in DWCSW). Probably as a result of the restricted range, our response variable is not outlier prone.

For IBI and the QHEI the available data are a single value per location. For other variables the data were reduced to a single value per location, using the median in case of multiple values for a location (a common practice in evaluation of water quality data).

Zinc and oxygen concentrations were transformed to logarithms. (pH is the log of a concentration – of hydrogen ions – by definition.) Logarithmic transformation of concentration measurements is conventional in regression analysis of water quality data. Each regressor was centered and scaled (after logarithmic transformation, for zinc and DO), to enhance comparability of coefficients among regressors.

The model was implemented with sign constraints on regression coefficients. A negative sign is expected for the coefficient for zinc, which at high concentrations is toxic to aquatic organisms. Low DO excludes some aquatic species and is often a consequence of eutrophication, for example from agricultural fertilizers. A well-known ecological effect of low pH (acidification) is to increase the availability of toxic metals such as zinc, cadmium, and aluminum. pH is in fact negatively correlated with zinc in our data. Preliminary analysis suggests that our sign constraints for regression coefficients are consistent with the data: In a

standard multiple regression of IBI on the four regressors, each coefficient had the sign in agreement with our constraint and the corresponding *t* statistic was greater than 3 in absolute value.

Additional constraints are a minimum count of 15 measurement locations per cluster (combining the location counts for all basins belonging to a cluster), and a minimum value of 0.1 for each mixing proportion. Except where otherwise indicated the results presented in this section are obtained with a 2-component model with the Dirichlet parameter α set equal to 50. Models with 3 and 4 components were also evaluated, and for the 2-component model we evaluated alternative values of α. For 3- and 4-component models, the clusters were not as well separated or interpretable as for the 2-component model. Further exploration of models with 3 or more components may reveal additional information at a scale of resolution finer than provided by the 2-component model. Results for the 2-component model suffice for present purposes.

The algorithm was programmed in R. Sampling of the multivariate normal distribution was accomplished using the R contributed package `mvtnorm` (Genz, Bretz, and Hothorn, 2004). Except where otherwise indicated the results displayed are based on a posterior sample of size 10000, excluding a burn-in sequence of 1000 cycles.

Figure 1 displays the result of an average linkage cluster analysis, using distances based on the ECCPs, performed with the base R function `hclust`(). (One minus ECCP is a distance measure satisfying the triangle inequality.) The dendrogram suggests well-defined clusters apart from the cluster assignments of a few basins.

We have not yet encountered indications of label switching within individual chains for the 2-component model, apparently because the components are well separated. However, the indices will sometimes be permuted between successive simulations. In order for the results to be comparable among simulations we impose an identification constraint on component

membership: From the ECCP matrix we observed that Basins 1 and 19 belong to different components with probability about one. Based on this observation we constrained basin 19 to belong to component 1 and basin 1 to belong to component 2. These constraints were implemented by setting membership probabilities $\hat{z}_{19,1}$ and $\hat{z}_{1,2}$ equal to one, and setting to zero the probabilities of belonging to other components, for basins 1 and 19. We suggest this type of approach to the label switching problem is applicable in clustering applications of FM models when the components are well separated, perhaps as indicated by a dendrogram based on ECCPs such as our Figure 1.

The identifying constraint could have relied upon some other set of basins. However, we observed that for each pair of basins that belong together with estimated probability 0.1 or less, one belongs to component 1 with probability at least 0.9, and the other to component 19, also with probability 0.9 or greater.

***** Figure 1 about here *****

Table 1 compares different values of the Dirichlet parameter α with respect to the posterior probability that a given basin belongs to Component 1. For most basins there is little uncertainty in the component assignment. The final row of the table gives the measure of uncertainty, averaged over basins. (For an explanation of the final column in Table 1, see remarks associated with Expression 3.) Based on Table 1, we can identify a set of basins that has relatively high certainty in assignment to components, and this identification is not sensitive to the value of α. For two basins with relatively greater uncertainty, the most probable assignment is observed to depend on α.

***** Table 1 about here *****

Figure 2 compares the two components with respect to the posterior distribution of each regression coefficient, including the intercept. There is a clear separation based on the intercept and the coefficient for QHEI, and somewhat less pronounced separation for the zinc coefficient.

We note that for each coefficient the posterior distribution is well-separated from zero, suggesting again that our sign constraints are not contradicted by the data.

***** Figure 2 about here *****

We used ArcView to evaluate the spatial disposition of measurement locations for the two clusters, relative to ecologically relevant features of the geography of Ohio. (ArcView is a trademark of Environmental Systems Research Institute, Inc.) Figure 3 displays the basins and component assignments. Also shown is the boundary between the drainages of Lake Erie and the Ohio River, a feature that may relate to the interpretation of the results (see discussion).

***** Figure 3 about here *****

A graphical check on convergence for the 2-component model was based on independent chains using the method of Gelman and Rubin (1992) as implemented in the R contributed package `coda` (Plummer, 2005). We applied the approach based on 5 chains with independent random initial partitions, to check convergence for the mixture log-likelihood, the mixing proportions, and the regression parameters. The results suggested that 1500 iterations suffice after discarding an initial burn-in sequence of 500 iterations. In addition numerous repetitions with $\alpha = 50$ during program development have provided no indications of local optima. A technique that we have found useful for searching for local optima is an index plot of the FM likelihood, combining in sequence the results for the independent chains (see for example Chipman *et al,* 2002).

5. Discussion

Our approach is in some ways an extension of the classification likelihood (CL) approach of Lipkovich *et al*. (to appear). In a finite mixture model, the Dirichlet prior for mixing proportions represents the highest level in the probability hierarchy. For a classification model the highest level in the probability hierarchy is the partition. For a Bayesian CL model, prior probabilities would be specified directly for partitions, for example, by assuming that each

partition is equally probable *a priori*. The CL approach of Lipkovich *et al*., like our approach, relies on a Metropolis algorithm to explore a space of partitions, allowing a distinct regression model for each cluster. Other features in common include the grouping of observations into CUs, with a requirement that measurement locations belong to the same component if they belong to the same CU, and a Metropolis proposal for updating the partition by reassigning a randomly selected CU to a randomly selected component. Our approach differs most importantly by providing inference jointly for the partition and parameters of the regression model. According to the approach of Lipkovich *et al*. partitions are compared based on the value of the optimized classification likelihood – optimized over regression parameters -- with a penalty of the BIC form (Schwarz, 1978). We have observed that when the Dirichlet parameter α is increased in value the result is in effect a CL algorithm, but without regression parameters eliminated by optimization or integration.

However we observe that when interest is in the partition, a significant simplification may be achieved by eliminating component distribution parameters, using exact or approximate Bayes factors. Such simplification may be helpful particularly in the case of a multivariate response.

Methods specifically used to identify regions may benefit by incorporation of some tendency for locations that are nearby in space to fall in the same cluster. An approach sometimes used is to include spatial coordinates among the variables included in a conventional cluster analysis. The approach may yield useful results but may be restrictive with regard to the geometry of region boundaries. Considering the correspondence between some cluster analysis procedures and modeling with multivariate normal mixtures (Gordon, 1999), the approach may favor clusters with elliptical boundaries. Lamon and Stowe (2002) used spatial coordinates in a Bayesian classification tree approach, resulting in regions with boundaries parallel to lines of latitude or longitude.

Our initial approach, applicable for a data set where measurement locations are classified by CU, is to require that locations belong to the same cluster if they belong to the same CU (Lipkovich *et al*., to appear). However, we think there may be multiple types of spatial constraint that deserve consideration. We observe that our explicit Metropolis step is potentially very flexible for purposes of implementing alternative spatial constraints. Modifications of the Metropolis proposal can be used to exclude certain partitions, while modifications of the acceptance criterion can be used to introduce a probabilistic preference for partitions with desirable properties.

The best correspondence we have found between our components and a geographic feature is a rough correspondence between our component 1 and the Lake Erie drainage area, and between our component 2 and the Ohio River drainage area (see Figure 3). However, this correspondence with drainage areas may be partly related to our requirement that locations belong to the same component if they belong to the same river basin.

Of considerable interest to environmental analysis is the degree of correspondence between results from a proposed method of cluster analysis and standard ecoregions. (Ecoregions are not shown in our Figure 3. Ecoregion maps are available from USEPA (2005).) The best correspondence between our results and a standard ecoregion is between our component one and the Huron and Erie Lake Plains (HELP) ecoregion. However, some river basins in the Lake Erie drainage have lowland stretches in the HELP and headwaters in the Eastern Corn Belt Plains (ECBP) ecoregion. With our approach, these parts cannot be assigned to separate components. If we ignore the headwater stretches of these basins, there is a better correspondence between our component 1 and the HELP.

McMahon *et al.* (2001) have suggested the use of statistical techniques for identifying regions of about the size of USEPA Level III ecoregions. Our results seem to provide a limited

corroboration of this suggestion. We find interpretable regions of about the size of Level III ecoregions, and with an imperfect correspondence of one component to a standard ecoregion.

Numerous simulations started from independent partitions indicate rapid mixing for our 2-component model as applied to the Ohio data, and no indications of local optima. This relatively favorable outcome may result from constraints on the signs of coefficients, on the sizes of clusters, and on ranges of mixing proportions. Indeed, when we relaxed the sign constraints for the regression coefficients a second optimum was evident in some simulations, having lower likelihood than the one reported here, and with the coefficient for zinc having the "wrong" sign. (While this observation is of some interest from the standpoint of the role of constraints, we feel confident that a model acceptable for use in predicting ecological quality should not have a negative coefficient for zinc concentration.)

Finally, we suggest that the application of statistical classification procedures for defining geographic regions useful for environmental analysis is a promising area for collaboration between statistical modelers and other scientists in view of the range of subject-matter and statistical issues.

References

Banfield, J.D., and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803-831.

Bensmail, H., Celeux, G., Raftery, A., and Robert, C. (1997). Inference in model-based clustering. *Statistics and Computing* **7**, 1-10.

Celeux, G., and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition 28:781-793. Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957-970.

Chipman, H.A., George, E.I., and McCulloch, R.E. (1998). Bayesian CART model search (with discussion*). Journal of the American Statistical Association* **95**, 935-960.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2002). Bayesian treed models. *Machine Learning* **48**, 299-320.

Denison, D.G.T., Mallick, B.K., and Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363-377.

DeSarbo, W.S., and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**, 249-282.

Diebolt, J., and Robert, C.P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society B* **56**, 363-375.

Dyer, S.D., White-Hull, C., Carr, G.J., Smith, E.P., and Wang, X. (1998). Bottom-up and top-down approaches to assess multiple stressors over large geographic areas. *Environmental Toxicology and Chemistry* **19**, 1066-1075.

Fraley, C., and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* **41**, 578-588.

Fraley, C., and Raftery, A.E. (2002a). Model-based clustering, discriminant analysis, density estimation. *Journal of the American Statistical Association* **97**, 611-631.

Gelman, A., and Rubin, D.B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-511.

Genz, A., Bretz, F., and Hothorn, T. 2004. The `mvtnorm` package. Manual available on the web site of the Comprehensive R archive network.

`http://cran.r-project.org/`

Gordon, A.D. (1999). *Classification*. 2nd edition, Chapman & Hall/CRC: New York.

Lamon, E.C., and Stow, C.A. (2004). Bayesian methods for regional-scale eutrophication models. *Water Research* **38**, 2764-2774.

Lipkovich, I., Smith, E.P., and Ye, K. (to appear). Detecting Pattern in Biological Stressor Response Relationships Using Model Based Cluster analysis. *Ecological and Environmental Statistics*.

McMahon, G., Gregonis, S.M., Waltman, S.W., Omernik, J.M., Thorson, T.D., Feeouf, J.A., Rorick, A.H., and Keys, J.E. (2001). Developing a spatial *framework of common ecological regions for the coterminous United States. Environmental* Management **28**, 293-316.

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

McLachlan, G., Peel, D., Basford, K.E., and Adams, P. (2000). The EMMIX software for the fitting of mixtures of normal and *t*-components. *Journal of Statistical Software* **4**: 1-14.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2005). The `coda` Package. Manual available on the web site of the Comprehensive R Archive Network

`http://cran.r-project.org/`

Robertson, D.M., and Saad, D.A. (2003). Environmental water quality zones for streams: a regional classification. *Environmental Management* **31**, 581-602.

Roeder, K., and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894-902.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B* **62**, 795-809.

Ter Braak, C.J.F., Hoijtink, H., Akkermans, W., and Verdonschot, P.F.M. (2003). Bayesian model-based analysis for predicting macrofaunal communities. *Ecological Modeling* **160**, 235-248.

USEPA Western Ecological Division. (2005).

`http://www.epa.gov/wed/pages/ecoregions.htm`

Viele, K., and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing* **12**, 315-330.

Wedel, M., and Kamakura, W. (2000). *Market Segmentation*. 2nd edition, Kluwer: Boston.

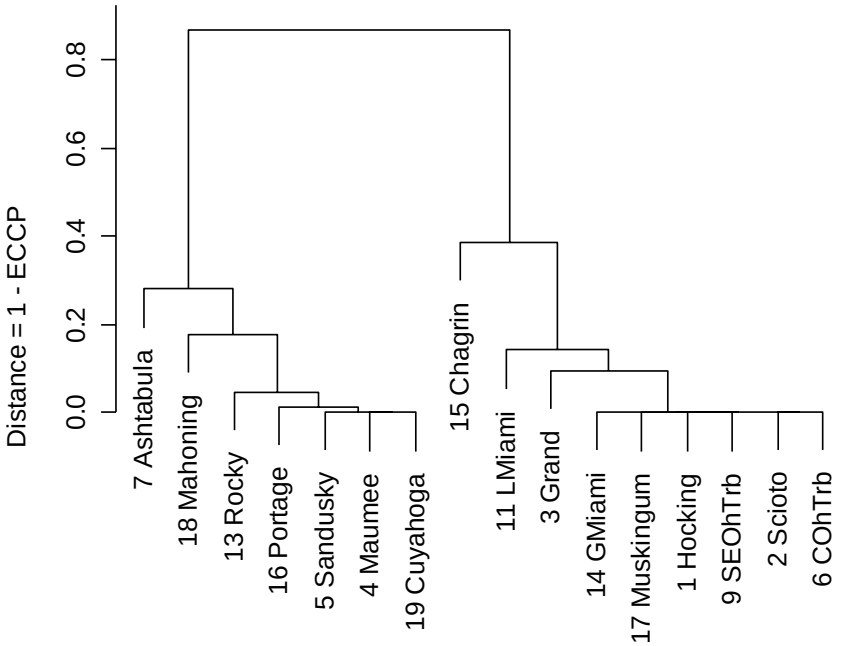

Figure 1. Dendrogram on use of Estimated Co-clustering Probabilities as a similarity.

Table 1. Estimated posterior probabilities of belonging to component 1 for each basin.[1]

Basin index [2]	Basin Name	Dirichlet parameter α		
		1	**50**	∞ [3]
1	Hocking R.	0.00	0.00	0.00
2	Scioto R.	0.00	0.00	0.00
3	Grand R.	0.03	0.07	0.07
4	Maumee R.	1.00	1.00	1.00
5	Sandusky R.	0.97	0.99	0.99
6	Centrial Ohio Tributaries	0.00	0.00	0.00
7	**Ashtabula R.**	**0.36**	**0.50**	**0.57**
9	SE Ohio Tributaries	0.00	0.00	0.00
11	Little Miami R.	0.04	0.05	0.08
13	Rocky R.	0.79	0.88	0.90
14	Greater Miami R.	0.00	0.00	0.00
15	**Chagrin R.**	**0.14**	**0.22**	**0.25**
16	Portage R.	0.88	0.96	0.96
17	Muskingum R.	0.00	0.00	0.00
18	**Mahoning R.**	**0.47**	**0.66**	**0.66**
19	Cuyahoga R.	1.00	1.00	1.00
Uncertainty averaged over basins		0.088	0.084	0.082

[1] For each model the posterior sample is obtained by combining 5 chains with independent initial partitions, each chain of length 1500-4500 after deletion of an initial burn-in sequence of length 500.

[2] Basins are indexed as in Dyer *et al*. (1998).

[3] The algorithm become insensitive to α at large values of that parameter. (See the remark following Expression (2).).

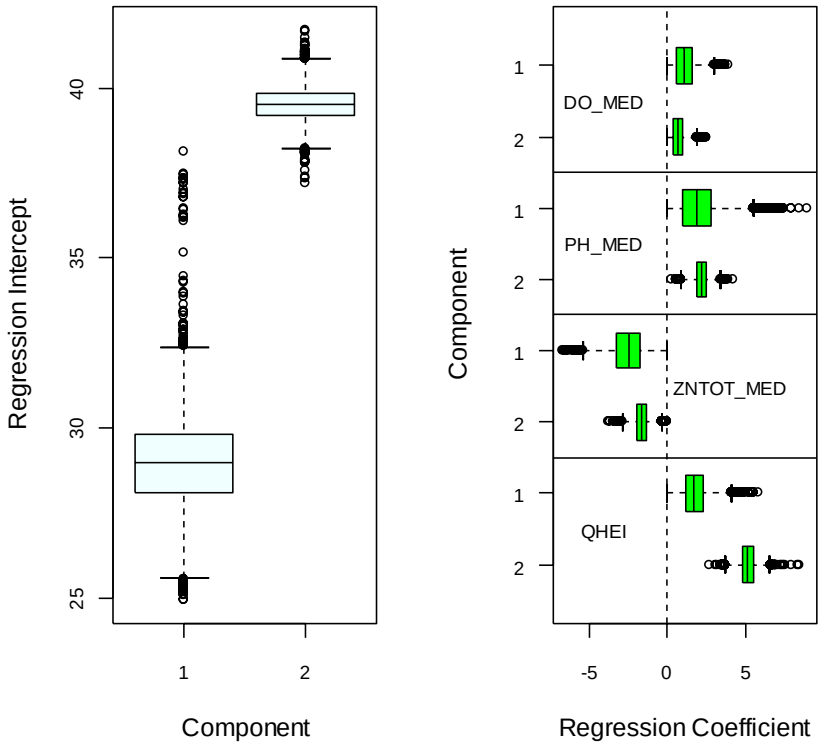

Figure 2. Distribution of posterior sample for component regression intercepts (left) and the coefficient for each regressor (right).

Figure 3. Spatial disposition of two clusters. The boundary between the Lake Erie drainage area and the Ohio River drainage area is indicated using a heavy line. River basins assigned to component 1 are shaded most lightly; those assigned to component 2 most darkly. Basins with intermediate shading did not contribute to the analysis because of missing values for some measurement locations.