ShiLi Miao · Susan Carstenn
Martha Nungesser
Editors

# Real World Ecology

## Large-Scale and Long-Term Case Studies and Methods

Foreword by Stephen R. Carpenter

Springer

*Editors*
ShiLi Miao
South Florida Water
Management District
West Palm Beach
FL, USA
smiao@sfwmd.gov

Susan Carstenn
Hawaii Pacific University
Kaneohe, HI
USA
scarstenn@hpu.edu

Martha Nungesser
South Florida Water
Management District
West Palm Beach
FL, USA
mnunges@sfwmd.gov

ISBN 978-0-387-77916-4 e-ISBN 978-0-387-77942-3 (Softcover)
ISBN 978-0-387-77941-6 e-ISBN 978-0-387-77942-3 (Hardcover)
DOI 10.1007/978-0-387-77942-3

Library of Congress Control Number: 2008941656



springer.com

# Chapter 5
# Bayesian Hierarchical/Multilevel Models for Inference and Prediction Using Cross-System Lake Data

**Craig A. Stow, E. Conrad Lamon, Song S. Qian, Patricia A. Soranno and Kenneth H. Reckhow**

**Abbreviations** AIC – Akaike's Information Criterion, ANOVA – Analysis of Variance, BCART – Bayesian classification and regression tree, BIC – Bayesian Information Criterion, CART – Classification and Regression Tree, DIC – Deviance Information Criterion, LIL – Log Integrated likelihood, MCMC – Markov Chain Monte Carlo, MLE – Maximum Likelihood Estimator, SBC – Schwarz's Bayesian criterion, TP – Total phosphorus

## 5.1 Introduction

Cross-system studies are commonly used for large-scale ecological inference (Cole et al. 1991). Many processes change slowly within a particular ecosystem, thus long time periods can be required to measure how changes in one process may influence changes in another. By using data from many systems researchers essentially substitute space for time, assuming commonality among the systems being compared. Comparing characteristics among systems helps researchers identify patterns that provide clues for understanding ecosystem function, generate testable hypotheses, and isolate cause–effect relationships.

In limnology, cross-system studies have been widely applied, in part because lakes are relatively discrete ecosystems with tangible boundaries, making their properties straightforward to evaluate and compare. Additionally, many important lake attributes, such as trophic status, can be well-approximated by a small number of quantitative measures. In the 1960s and 1970s cross-system lake comparisons were influential in resolving the limiting nutrient debate. The work of Richard Vollenweider, in particular, popularized this approach, and led to nutrient loading concepts that are still widely applied in aquatic ecosystem management (Vollenweider 1968, 1969, 1975, 1976).

C.A. Stow (✉)
NOAA Great Lakes Environmental Research Laboratory, 2205 Commonwealth Blvd., Ann Arbor, MI 48105, USA
e-mail: craig.stow@noaa.gov

S. Miao et al. (eds.), *Real World Ecology*, DOI 10.1007/978-0-387-77942-3_5,

Early efforts to quantify relationships among lakes may have been largely ad hoc; electronic calculators were not widely available until the mid-1970s and access to fast computers was limited. Thus, even relatively simple models like linear regression and ANOVA were tedious to compute for more than a small number of observations. With the increasing availability of fast, inexpensive computing power researchers began to use cross-system lake data to develop mathematical relationships useful for quantifying limnological processes and to build models for decision-making (Reckhow and Chapra 1983).

The use of cross-system data for lake-model development has also been fostered by a general lack of extensive long-term data from individual lakes. However, individual lake behavior can be idiosyncratic – highly dependent on features such as the landscape setting in which the lake is located (Stow et al. 1998). Thus, models developed from cross-system data may not accurately capture the behavior of a particular lake. Because management decisions are typically made on a lake-by-lake basis, it is arguably preferable to construct lake-specific models to support management actions such as Total Maximum Daily Loads (National Research Council 2001). Nevertheless, with some exceptions (Stow et al. 1997, Lathrop et al. 1998), relatively few lakes have been sufficiently well-monitored to provide adequate data for individual model development. This is particularly true in lake-rich states in the US and other regions of the world where limited resources prohibit intensive data collection on more than a few lakes.

Additionally, data representing a range of conditions are necessary to assess the functional form of a model and estimate model parameters. Individual lakes in approximately steady-state conditions are unlikely to generate data of sufficient variance for accurate model estimation even over long time periods. In such cases extrapolation beyond the observed conditions may be required to predict the likely changes that will occur under alternative management actions, resulting in considerable predictive uncertainty. Thus, even if many years of data are available for a particular lake, augmenting the lake-specific data with data from other lakes can increase predictive accuracy for the lake under consideration.

Building models for prediction from cross-system data is based on the supposition that the relationship between the response and predictor variables is the same for all the lakes used to estimate the model. While this presumption is never exactly correct, most modelers implicitly hope that it is close enough to the truth to make useful inferences and decisions. A popular strategy to increase the likelihood of similar behavior is to group lakes based on common attributes. Features used to categorize lakes have included lake type (natural vs. reservoir), geographic setting, and geomorphology (Canfield and Bachman 1981, Reckhow 1988, Malve and Qian 2006) and models based on such groupings are then used for individual lake forecasts (Hession et al. 1995). The goal of this strategy is to "borrow" information from other similar lakes to increase the accuracy of prediction for a particular lake, so that the risk of making a bad decision is acceptably low.

But are models developed from a cross-section of lakes sufficiently accurate to make good management decisions for an individual lake? Moreover, how uncertain can a model be and still be useful for effective management decision-making? The answer to these questions is context specific, depending on the stakes associated with the pending decisions. If the consequences of a poor decision are not very severe, high model uncertainty may not matter much. Alternatively, in high-stakes decision situations, models with large uncertainties may not differentiate the likely outcomes of one management alternative from another, providing little basis for selecting a management option.

Given a choice of models, a decision-maker is likely to use the model with the lowest uncertainty. A model that provided correct forecasts 98% of the time would be a clear choice over one that was accurate only 75% of the time. With 98% accuracy, management actions could be chosen based only on the societal value of the consequences of those actions. A model with higher uncertainty may still be informative, but applying this model requires decision-makers to hedge their decisions by considering a range of possible outcomes, based on knowing how uncertain the model is. Therefore, while uncertainty may arise from a lack of knowledge, quantified uncertainty provides information that is useful for both picking the best model and applying that model for decision support.

Often, however, decision-makers are given models or model results with no supporting information regarding uncertainty – providing an illusion of precision (Pappenberger and Beven 2006). How is this information useful for high-stakes decisions?

Uncertainty is typically expressed as a probability statement. The term probability has several interpretations; the most widely appreciated is the long-term relative frequency of a particular event. Coin-flipping is a common example of this probability notion. The probability of flipping a coin and obtaining "heads" is approximately 50% – meaning that if a coin is flipped enough times approximately 50% of those flips will result in the coin landing heads-side up. Another interpretation of probability is that it represents a degree-of-belief (Winkler 2003). If an individual is presented with a coin and asked the probability that the coin will land heads-side up when flipped, it is likely that the response will be: "approximately 50%". In this instance there is no long-term relative frequency involved, in fact the coin has not been flipped even once yet. The response "50%" represents the respondent's degree-of-belief that the coin will land heads-side up. This belief is based on prior experience resulting in knowledge of what usually happens when a coin is flipped. If the respondent had reason to believe that they were being tricked with a coin that had been engineered to produce heads they might indicate 90 or 95%. Thus, the degree-of-belief notion of probability is a quantification of the confidence an individual has in the occurrence of a particular result.

These two notions of probability are not mutually exclusive; in fact they can be reconciled using Bayes theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\theta \pi(\theta)f(y|\theta)\mathrm{d}\theta} \tag{5.1}$$

The interpretation of Bayes theorem is that prior beliefs (those held before the experiment), represented by $\pi(\theta)$, are combined with new information from the experiment, contained in the likelihood function, which is represented by $f(y \mid \theta)$, to obtain posterior beliefs, represented by $\pi(\theta \mid y)$. For most practical applications, the denominator on the right side of the equation can be regarded as a scaling constant and essentially ignored.

If, for example, before flipping a coin we believed that the probability of heads was 50%, and in our experiment of 100 flips we obtained 50 heads and 50 tails our belief would be unchanged; the posterior belief would be the same as the prior belief. If, however, the coin-flipping experiment resulted in 95 heads it is likely that our posterior belief in the probability of heads would move toward 95%, depending on whether or not we thought the 95 head outcome was typical or accidental.

While Bayesian approaches have been known in ecology and the environmental sciences for some time (Reckhow 1990), they are being increasingly used (Ellison 2004), in part because they have only become practically feasible since the widespread availability of fast, inexpensive computers. Many models are analytically intractable in a Bayesian context, particularly nonlinear models or models with more than just a few parameters to be estimated. Using a Bayesian approach in such cases was virtually impossible until recently. But the advent of cheap computing has fostered the development of algorithms that provide precise numerical approximations for most problems, making the routine application of Bayes theorem a practical option.

Bayesian approaches are not new; their use predates what are often referred to as "frequentist" or "classical" methods, which were developed in the early-mid 1900s (Salsburg 2001). Classical approaches are what most ecologists have been trained in since the 1960s, and usually involve setting up a null hypothesis, then performing a "significance test" to evaluate the validity of the null hypothesis. Many textbooks present classical methods as a coherently conceived approach to scientific inference, and their application has become deeply engrained in ecology. Thus, many ecologists perceive them to be inviolate rules that govern the way science is properly conducted, and finding "significant" results has become an end in itself rather than a means to an end. In fact, classical significance testing arose from a somewhat rocky fusion of two contrasting schools of thought, contributing to considerable confusion among applied scientists when interpreting their results, as *p*-values and significance levels are often interpreted to be synonyms (Hubbard and Bayarri 2003, Hubbard and Armstrong 2006).

Working in a Bayesian framework offers distinct advantages over a classical null hypothesis testing framework. A Bayesian approach provides a posterior probability distribution for the model parameters, $\pi(\theta \mid y)$, which indicates the

relative probabilities of all possible values of $\theta$, conditional on the observed data, $y$. This posterior distribution can be used to support a wide range of decisions applying many possible decision criteria. Predictions for $y$ are made by evaluating the model over the entire posterior parameter distribution, resulting in a predictive probability distribution for future $y$ values.

Classical null hypothesis testing approaches are much more constrained. The decision to be made is to either accept or reject an *a priori* null hypothesis, and the decision criterion is chosen to minimize the chance of accepting a false null hypothesis. Acceptance or rejection of the null hypothesis is based on a $p$-value, which is the probability of obtaining results at least as extreme as the observed data, conditional on the null hypothesis being true. Many researchers find $p$-values (symbolically expressed as $\pi(\geq y \mid \theta)$ where $\theta$ is typically zero) confusing, often misinterpreting them to convey the same information as a posterior distribution.

Bayesian approaches are sometimes disparaged as subjective because of the need to provide prior information regarding parameter values (Dennis 1996), but in fact modelers often fix parameters at specific values based on precedent, published literature values, or expert-judgment, and this practice is rarely criticized. Bayes theorem offers modelers considerably more flexibility. If a parameter value is well-known then a fixed value can be used, but if the modeler has very little prior information regarding a parameter value, or wants the posterior distribution to be determined by the data, then a non-informative or vague prior distribution can be chosen. Non-informative priors typically have a large variance, thus they convey minimal information about the value of a model parameter. Alternatively, semi-informative priors, with intermediate variances, can be used for parameters that are not known precisely, but are likely to be within a limited range (Stow and Scavia 2008). In any case, prior specification in Bayes theorem is explicit, allowing other researchers the opportunity to evaluate the sensitivity of model-based inferences to the choice of the prior.

## 5.2 Multilevel/Hierarchical Models

Although lake management decisions are usually made on a lake-by-lake basis, the scientific information for such decisions is often based on cross-system data. On one hand, a model developed using data from multiple lakes will almost certainly be less accurate for a specific lake, because the model represents the average of the lakes. On the other hand, a model based only on lake-specific data will have a large uncertainty because of the small lake-specific sample size. Multilevel models provide a rigorous framework to systematically combine information from several sources and appropriately weight the group-specific or lake-specific information depending on the degree of similarity to other groups or lakes in the data set.

In a multilevel model, parameters representing individual lakes, or groups of lakes, are given a probability structure. For example, within a lake, observations of a response variable, such as chlorophyll *a* concentration, exhibit a lake-specific mean, and a certain level of within-lake variability. If the within-lake variability is expected to be similar for lakes belonging to the same group, then the between-lake variability can be expressed in terms of the variability among the lake-specific means. This variability among lake-specific means can be modeled with a normal (or other) distribution and the parameters of this distribution can be estimated from the cross-system data. In some cases, observations may be nested – there may be distinct basins within lakes, lakes within similar groups, groups within a region, etc., and the probability model capturing each of these levels of variability assumes a hierarchical configuration.

To explain the difference between the multilevel approach and more traditional approaches, we use a simple example of estimating lake-specific means of a response variable. Mathematically, we would model this between and within-lake variability using a model similar to an ANOVA model

$$y_{ij} \sim N(\mu_i, \sigma^2), \tag{5.2}$$

where $y_{ij}$ is the $j$th observation from the $i$th lake, $\mu_i$ is the mean of the $i$th lake and $\sigma^2$ is the variance of the observations. When individual lakes are modeled separately, each $\mu_i$ is estimated to be a lake-specific mean. When cross-system data are used, all lakes implicitly share the same mean. The multilevel approach is a compromise between these two extremes. It treats each individual lake as a separate entity, but the lake-specific means are structured to share the same prior distribution:

$$\mu_i \sim \mathrm{N}(\mu, \tau^2), \tag{5.3}$$

where $\mu$ and $\tau^2$ are the mean and variance, respectively, of the $\mu_i$. By using a shared prior distribution of all lake-specific means, the multilevel model results in at least two improvements over more conventional approaches: first, the shared prior distribution provides a connection between individual lakes. Because of the common prior, the lake specific means are now estimated as a weighted average of individual lake mean and the overall mean.

$$\bar{\mu}_i = \frac{\left(n_i/\sigma^2\right)\bar{y}_i + \left(1/\tau^2\right)\bar{y}_{\mathrm{all}}}{\left(n_i/\sigma^2\right) + \left(1/\tau^2\right)} \tag{5.4}$$

The weights are determined by individual lake sample size ($n_i$), individual lake variance ($\sigma^2$), and the variance among lake-specific means ($\tau^2$). When the sample size of a lake is large, the multilevel estimated lake specific mean will be close to the sample mean. When the sample size is small, the multilevel estimated lake specific mean will be close to the overall mean. If there are no

data from a lake, the multilevel estimate of the lake specific mean is the overall mean. In addition to the sample size, the multilevel estimate also considers the levels of between and within lake variances. If the between-lake variance ($\tau^2$) is high, the weight on overall mean will be small, individual lake mean will be weighted heavier, and vice versa. This partial pooling of the cross-system data allows system specific estimates without requiring a large sample size for each system. Second, by pooling data from multiple lakes, the uncertainty associated with lake-specific prediction is reduced. If we refer to the cross-system approach, where data from multiple lakes are pooled, as complete pooling of all lakes, and refer to the separate analysis of individual lake approach as no pooling, the multilevel modeling approach is called partial pooling of the data. By partial pooling, we balance the information from individual lakes and the overall average from all lakes. The individual lake data are more variable (more uncertain) while the overall average is less variable (more certain).

Compared to the separate lake means, the multilevel-based estimates of individual lake means are pulled towards the overall mean. Lakes with larger sample sizes are pulled less, while lakes with smaller sample sizes will be pulled more to the overall mean. This pulling (also known as Bayesian shrinkage) represents a form of information discounting. If a lake with few observations, or a large variance, has a mean chlorophyll *a* concentration that is much higher than the overall mean of all the lakes, we are likely to doubt the validity of the anomalously high estimate. The multilevel estimate provides a sensible way of discounting the information that is less trustworthy. This is consistent with the way that Bayes theorem pools information from different sources. The information represented in the overall mean can be seen as the prior for an individual lake, while data from the lakes are treated as observations. Under the hierarchical framework, prior distributions are typically needed for the model parameters that are of secondary interest, thus their influence on the parameters of primary interest is indirect and often minimal. In our case, we need to supply prior distributions for the mean of group means of model coefficients and the within and between group standard deviations of model coefficients. When there are many groups, the mean of group means is of less interest and we usually do not have much information on the within and between group standard deviations of model coefficients. As a result, non-informative priors are used.

The multilevel approach can be readily extended to more complex models where the variable of interest is a linear or nonlinear function of one or more predictor variables, with unknown model parameters. In our example, lake chlorophyll *a* concentration is the variable of interest and we model the log of chlorophyll *a* as a simple linear function of the log of total phosphorus concentration. The multilevel model partially pools the group data by introducing common prior distributions for model coefficients. With this approach, the model structure is:

$$\log(\text{Chla})_{ik} \sim \text{N}(\mu_{ik}, \tau^2) \tag{5.5a}$$

and

$$\mu_{ik} = \alpha_k + \beta_k \log(\mathrm{TP}_{ik}), \tag{5.5b}$$

$$a_k \sim N(\alpha, \sigma_\alpha^2), \tag{5.5c}$$

$$\beta_k \sim \mathrm{N}(\beta, \sigma_\beta^2), \tag{5.5d}$$

where $\log(\mathrm{Chla})_{ik}$ is the natural log of chlorophyll *a* concentration from lake *i* in group *k*, $\mu_{ik}$ is the mean of the *i*th lake in group *k*, $\tau^2$ is the variance of log(Chla) at a given value of log(TP) (assumed to be the same for all lakes), $\alpha_k$ and $\beta_k$ are the intercept and slope parameters, respectively, for group *k*, $\boldsymbol{\alpha}$ is the mean of the $\alpha_k$s, $\boldsymbol{\beta}$ is the mean of the $\beta_k$s, and ${\sigma^2}_\alpha$ and ${\sigma^2}_\beta$ are the respective variances of the $\alpha_k$s and $\beta_k$s.

Computing a multilevel model can be done in either a Bayesian or non-Bayesian context. The non-Bayesian implementation is usually referred to as a random, fixed, or mixed-effect model. Typical random, fixed, and mixed-effect models use a maximum likelihood estimator (MLE) producing a result similar to a Bayesian model using vague prior distributions for $\sigma^2$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, ${\sigma^2}_\alpha$, and ${\sigma^2}_\beta$. However, estimating the variances can be difficult using MLE when the number of lakes or number of groups is small (single digit). The Bayesian method is more flexible, especially when proper prior information is available. A Bayesian implementation of multilevel modeling is typically calculated using Markov chain Monte Carlo (MCMC) simulation (Qian et al. 2003) and can readily be programmed in WinBUGS (Lunn et al. 2000) which is a free downloadable software. Multilevel/hierarchical models are similar in concept to random coefficient and empirical Bayes models which have been previously demonstrated with cross-system lake data (Reckhow 1993, Reckhow 1996). They can be very useful for organizing ecological data and have been used to synthesize information in a cross-system data set of Finnish lakes (Malve and Qian 2006).

## 5.3 Finding Groups in Data

As noted earlier, lake water quality models may be classified by the level of pooling used in the estimation of their parameters, and have historically either been developed in a lake-specific context (no pooling) or from a cross-system study (complete pooling). The simple linear structure of a regression model does not always apply to an entire cross-system dataset in the complete pooling situation. There are many reasons that this may be true for lake water quality data, including factors such as regional and local differences in climate, geology, morphometry, land use, land cover, and food web structure. An open question with regard to partially pooled lake models is "How do we determine which of the above mentioned factors to use to create lake groups for pooling?"

In some situations, a natural hierarchical structure is present in the data and may be exploited for the purpose of assigning lakes to groups for pooling. Lakes occur in watersheds, watersheds in ecoregions, etc. These are categorical or factor variables, in which observations take multiple discrete values or levels with in each factor. In such instances, it may be of interest to determine if one or more levels of the factor may be combined to reduce model complexity. Further, there may be thresholds, or change points, along the axes of continuous variables that may be used to define groups (above/below threshold) for pooling.

Chipman et al. (2002) provide an algorithm to obtain models that may better describe simple structure in cross-system data, by sub-setting the data and then fitting separate sub-models for each subset. We used Bayesian Treed models (Chipman et al. 2002) to determine if model predictive performance could be improved by fitting our simple linear regression model to subsets of cross-system lake data. Bayesian CART and Bayesian Treed models are both enhancements to the more familiar Classification and Regression Tree (CART) procedure (Breiman et al. 1984). Bayesian Treed Models select subsets of observations on partitions of the matrix of predictor variables for which linear models are estimated. Further, they identify subsets of the predictor variables for which linear model performance is improved as measured by predictive log integrated likelihood (LIL). Chipman et al. (2002) provide full details regarding Bayesian Treed models.

Use of Bayesian CART (Chipman et al. 1998) and Bayesian Treed models (Chipman et al. 2002) allows predictors to work together in a nonlinear, non-additive fashion by virtue of their tree-based structure, because the disjoint partitions on the predictor space provide for global non-linearity and non-additivity, even though model structure may be locally linear (and additive). We begin our discussion of methods with a general discussion of tree based methods, followed by a more specific description of Bayesian CART and Bayesian Treed models.

### *5.3.1 Tree-Based Models*

Tree-based models are useful for classification and regression problems in which the analyst cannot (or does not want to) specify *a priori,* the form of important interactions between explanatory (or independent) variables (Clark and Pregibon 1992). Tree models are easy to interpret, invariant to monotone transformations of the predictors, and able to capture interaction effects among the independent variables (non-additivity). Tree-based models made their first appearance in the statistical literature due to Sonquist and Morgan (1964). Much of their recent development is due to the work of Breiman et al. (1984). Because they are a computationally intensive procedure, their use has grown concurrently with the advent of the personal computer. Clark and Pregibon (1992) provide a description of classification and regression trees, along with

examples. Classification and regression trees have been used in numerous ecological studies (e.g., Magnuson et al. 1998, Lamon and Stow 1999, Qian and Anderson 1999). Bayesian Treed Models have been used with cross sectional data to link nutrients to chlorophyll *a* concentrations in lakes of the continental U.S. (Lamon and Stow 2004, Freeman et al. 2008), and in Finnish lakes (Lamon et al. 2008).

To fit a regression tree, the algorithm begins with the root (or parent) node. The root node contains all the observations and their associated variability. The data are split by binary recursive partitioning into increasingly homogeneous subsets until within-node variability is below some user-specified value. A terminal node is one that cannot be split further according to the user specified rules. Terminal nodes are also called leaves, consistent with the tree analogy. For every split made, the algorithm uses all unique values along the axes of each predictor variable as candidate values for splitting the dataset.

The process starts by defining the deviance of the root node (all of the data) as:

$$D(\mu) = \sum (\mathrm{y_i} - \mu)^2 \tag{5.6}$$

where $y_i$ are the observations within the node and $\mu$ is the node mean. Then each candidate predictor variable is examined to find a point that splits the response variable into two new nodes, a left and right, where:

$$D(\mu_L) = \sum (\mathrm{y_i} - \mu_L)^2 \tag{5.7a}$$

and

$$D(\mu_R) = \sum (\mathrm{y_i} - \mu_R)^2 \tag{5.7b}$$

are the deviances of the left ($\mu_\mathrm{L}$) and right ($\mu_\mathrm{R}$) nodes. The split that maximizes the deviance reduction, defined as:

$$\Delta D(\mu) = D(\mu) - \{D(\mu_L) + D(\mu_R)\} \tag{5.8}$$

is chosen, and the process begins again at the left and right nodes. The result is analogous to a dichotomous key where successive choices are made regarding the value of the response variable, based on predictor characteristics.

A comparison to ANOVA provides some qualitative insight into the CART procedure. ANOVA attempts to answer the question "Is the mean response different among the various levels of the chosen factors?" (Chambers et al. 1992). CART searches for the levels of the factors (as defined by splits) that have different means, and does not restrict the search to additive, globally linear models. When there is just one predictor variable, tree models are step functions, and with two predictors the partitions of a tree model may be plotted on a bivariate plot of the pair of predictor variables.

Some limitations of the method include the fact that it results in discontinuities at the partition boundaries rather than providing smooth transitions between partitions, as well as the inability to provide good approximations to linear and additive functions. Uncertainty regarding selection of the predictor variable upon which to partition the data, as well as the value at which to make the split once the variable is chosen, is dealt with in CART models only one node at a time. This decision is based on maximizing $\Delta D$. Because $\Delta D$ is calculated considering only the current node and the resulting daughter nodes, there may be a split of the current node that yields a less than maximum $\Delta D$ in the two daughter nodes, but provides a much better (i.e., larger) $\Delta D$ in subsequent splits.

#### 5.3.1.1 Bayesian CART and Bayesian Treed Models

These limitations are addressed using the Bayesian approach to the classification and regression tree algorithm (Chipman et al. 1998, Chipman et al. 2002). Instead of the binary recursive partitioning approach of conventional CART models, Bayesian CART (BCART) uses MCMC methods to explore the tree structure. The MCMC approach is computationally more demanding than the recursive partitioning used by conventional CART, but produces trees of lower overall deviance.

The difference between BCART and Bayesian Treed models is the specification of the end node model. BCART (and conventional CART) uses the mean of $y$ value in each leaf as the end node model, while a Bayesian Treed model uses a simple linear regression. By using a richer structure on the terminal nodes, we transfer model structure from the tree to the terminal nodes. We therefore expect smaller, more interpretable trees to result.

It is difficult to find the "best" tree using any tree-based method. The conventional approach to this problem is to use a "greedy" algorithm to "grow" a tree, then "prune" it back to avoid overfitting, as in conventional CART. These greedy algorithms usually grow a tree by sequentially choosing splitting rules for nodes based on maximizing some fitting criterion (Chipman et al. 1998). This approach produces a sequence of trees, all of which are refinements of the previous tree in the sequence severely limiting the exploration of all possible trees. In contrast, the Bayesian approach to CART consists of a prior specification and stochastic search (Chipman et al. 1998), exploring a much richer set of candidate trees. The Bayesian Treed models presented here were fit using software available by download free of charge: (http://faculty.chicagogsb.edu/robert.mcculloch/research/code/CART/index.html).

## 5.4 Comparing Models

Several criteria have been proposed for use in model comparison and selection. Many proposed criteria have a component that quantifies goodness of model fit, along with a component that penalizes model complexity. Among these

criteria are the Akaike Information Criterion (AIC, Akaike 1973), the Bayesian Information Criteria (BIC, Schwarz 1978), and the Deviance Information Criterion (DIC, Spiegelhalter et al. 2002). The goodness of fit component for all these make use of the deviance, $D(y,\theta) = -2\log[p(y|\theta)]$, where $y$ is response data associated with a parameterized model, $\theta$ are the parameters of that model, and $p(y|\theta)$ is the likelihood. Note that the likelihood is associated with a specific set of data, $y$. For this reason, criteria that are based on deviance (and therefore, on the likelihood) should be based on the same data $y$ for all models considered.

The Akaike Information Criterion (AIC) is calculated with the formula

$$\mathrm{AIC} = D(y,\theta) + 2p, \tag{5.9}$$

where $p$ represents the number of parameters in the fitted model. Note that the penalty, $2p$, increases with increasing complexity (number of model parameters), and a model with a lower AIC is preferred over a model with a larger AIC.

The Bayesian Information Criterion, also known as Schwarz's Bayesian criterion (SBC), is calculated according to the formula

$$\mathrm{BIC} = D(y,\theta) + p \times \log(n), \tag{5.10}$$

where $p$ again represents the number of parameters and $n$ the number of observations in the fitted model. As with AIC, a model with a lower BIC is preferred over a model with a larger BIC. Functions for calculation of AIC and BIC are readily available in the R statistical graphics package.

In the context of a Bayesian hierarchical model, the number of independent parameters included in the model is difficult to determine, which makes the use of AIC or BIC problematic. DIC has been proposed for model comparison in this context. Spiegelhalter et al. (2002) show that the effective number of parameters in a complex hierarchical model, $p_D$, can be computed as

$$p_D = \overline{D(y,\theta)} - D(y,\bar{\theta}). \tag{5.11}$$

The first term on the right is the mean deviance of the model using all of our estimates (samples) of $\theta$, and the second term on the right is the deviance of the model using only the mean value of our estimate of $\theta$ (mean of our samples). The deviance information criterion is then

$$\mathrm{DIC} = \overline{D(y,\theta)} + p_D. \tag{5.12}$$

As with the other criteria, a lower value of DIC is preferred over a higher value. Spiegelhalter et al. (2002) offer guidelines for using DIC to compare competing models similar to those suggested by Burnham and Anderson (1998) for

interpretation of differences in AIC between models, such that models within 1–2 of the "best" deserve further consideration and 3–7 have considerably less support. Calculation of DIC is relatively simple in MCMC, and a function for calculating DIC is implemented in WinBUGS.

## 5.5 Our Analyses

The goal of our analysis is to illustrate the utility of Bayesian and multilevel approaches for constructing models using cross-system data. This particular model is developed to help managers assess target phosphorus concentrations in lakes where data may be sparse. We used data from 382 Michigan lakes and reservoirs ($\geq 0.20$ km$^2$), with public access, sampled by the MI-Department of Environmental Quality (Fig. 5.1). Water data were collected from each lake on a single date during the summer stratified season (July–September) for chlorophyll *a*, Secchi depth, total phosphorus, total nitrogen, alkalinity, chloride, dissolved silica, and true water color. We separated the lakes into two types: natural lakes (lakes with or without a water control structure on it, but with little change in water level), and reservoirs (lakes that were created by damming of a river with a range in water residence time of several days to years). We also characterized each lake by surface area, shoreline development factor (ratio of shoreline perimeter to the perimeter of a circle of area equal to the lake), average depth, and catchment area defined as the cumulative catchment (including all connected upstream lakes and streams). Finally, we included ecoregion


**Fig. 5.1** Study area with locations of the 382 Michigan lakes

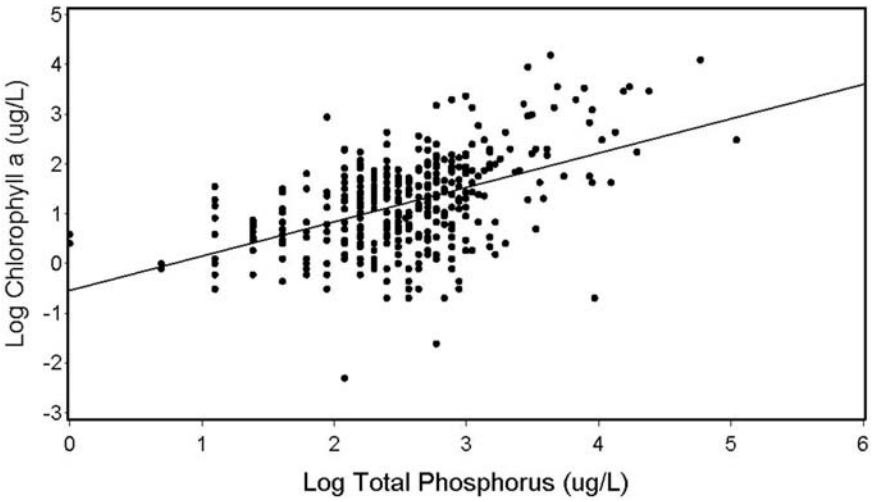



**Fig. 5.2** Log Chlorophyll *a* as a function of log Total Phosphorus for the 382 Michigan lakes

classifications including Omernik Level III ecoregion (Rohm et al. 2002), Great Lakes Basin, and landscape position (Martin and Soranno 2006).

Overall, the data exhibit an approximately linear relationship between log of chlorophyll *a* concentration and log of total phosphorus concentration, though there is considerable scatter about a simple regression line fit to the data (Fig. 5.2). This residual scatter will lead to increased uncertainty in the model parameter estimates and predictions, thus, it may be advantageous to subset the data into logical groups that reduce the scatter. We present the results from three models: a completely pooled model with no groups; a model with four groups based on Omerniks ecoregion; and a model with two groups based on the results from a Bayesian Treed analysis.

### *5.5.1 Completely Pooled Model*

Our first model, using complete data pooling, is analogous to the simple linear regression model:

$$\log(\text{chla}) = \alpha + \beta \log(\text{TP}) + \varepsilon \tag{5.13}$$

$$\varepsilon \sim N(0, \sigma^2) \tag{5.13a}$$

where log(chla) is the natural log of the chlorophyll *a* concentration, log(TP) is the natural log of the total phosphorus concentration, $\alpha$ and $\beta$ are the intercept and slope parameters, respectively, each to be estimated from the data, and $\varepsilon$ is

an additive model error term that is normally distributed with mean = 0 and variance = $\sigma^2$ where $\sigma^2$ is also estimated from the data.

To estimate this model using Bayes theorem requires specification of three prior distributions, one for each of the model parameters, $\alpha$, $\beta$, and $\sigma^2$. For this example, we programmed WinBUGS using non-informative priors so that the results would not be influenced by information outside of the data. For both $\alpha$ and $\beta$, we used a normal prior distribution with mean = 0 and variance = 10,000, and for $1/\sigma^2$ we used a gamma prior with scale and shape parameters both = 0.001. Variances in WinBUGS are specified by their inverses (hence the use of $1/\sigma^2$ which is referred to as the precision) and gamma distributions are a common non-informative prior for the precision. This choice has some caveats as we will illustrate in our second example model.

Because we used non-informative priors, the resulting posterior densities for $\alpha$ and $\beta$ (Fig. 5.3) convey information closely analogous to what would be obtained using a classical approach for a simple linear regression model. The important difference, however, is that rather than confidence intervals, which cannot be used to infer the probability of any particular result, we can make intuitive probability statements from these posterior distributions. Summary statistics for $\alpha$, for example, indicate that the mean and median of $\alpha$'s posterior are both –0.55 and the standard deviation of the posterior is 0.16 (Table 5.1). The 2.5 and 97.5 percentiles indicate that there is a 95% probability, given this model and these data, that $\alpha$ has a value between −0.86 and −0.23. Similar inference can be made for $\beta$, and $\sigma$, as well as for predictions of chlorophyll *a* in both the log and natural metrics.

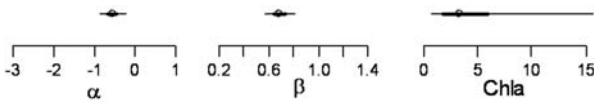



**Fig. 5.3** Parameter estimates and Chlorophyll predictive distribution for the fully pooled model. The circle represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

**Table 5.1** Summary statistics for completely pooled model

| Quantity | mean | median | sd | 2.5% | 97.5% |
|---|---|---|---|---|---|
| $\alpha$ | −0.55 | −0.55 | 0.16 | −0.86 | −0.23 |
| $\beta$ | 0.69 | 0.69 | 0.06 | 0.57 | 0.81 |
| $\sigma$ | 0.82 | 0.81 | 0.03 | 0.76 | 0.88 |
| Log predicted chlorophyll *a* | 1.17 | 1.17 | 0.82 | −0.42 | 2.79 |
| Predicted chlorophyll *a* | 4.50 | 3.23 | 4.4 | 0.66 | 16.3 |
| DIC | 930.2 | | | | |

Figure 5.3 also depicts a predictive distribution for chlorophyll *a* at a log TP value of 2.5 (the mean value of the data – chosen for convenience). At this log TP value chlorophyll *a* has 95% probability of being between 0.66 and 16.3 μg/L (Table 5.1). The distribution exhibits a right skew and median and mean values of 3.23 and 4.50 μg/L, respectively, illustrating the separation of the mean and median that occurs when models estimated under a log-transformation are retransformed back to the natural metric (Stow et al. 2006).

### 5.5.2 Ecoregion Model

Our second model has four groups based on Omernik's ecoregion (Rohm et al. 2002). Ecoregion might be considered a reasonable *a priori* basis for assigning groups, assuming that the ecoregion designation implies features consistent with distinctions in lake behavior. In this case, the model is

$$\log(\text{chla}) = \alpha_j + \beta_j \log(\text{TP}) + \varepsilon \tag{5.14}$$

$$\varepsilon \sim \text{N}(0, \sigma^2) \tag{5.14a}$$

where $j = 1$–4 with 175, 27, 179, and only 1 lake in groups 1–4, respectively (note: 1–4 correspond to ecoregions 50, 51, 56, and 55, respectively). A hierarchical structure results by imposing normal prior distributions on the $\alpha_j$ and $\beta_j$ with respective means $\mu_\alpha$ and $\mu_\beta$ and respective variances $\sigma^2\alpha$ and $\sigma^2\beta$. Both $\mu_\alpha$ and $\mu_\beta$ were assigned non-informative priors, normal distributions with mean = 0 and variance = 10,000, while the priors for $1/\sigma^2{}_\alpha$ and $1/\sigma^2{}_\beta$ were gamma with both scale and shape parameters = 0.001. Note that though each ecoregion has a distinct intercept and slope, a common variance, $\sigma^2$, was assigned. We also used a gamma prior with scale and shape parameters = 0.001 for $1/\sigma^2$. Additionally, to remove the correlation between the intercepts and slopes we centered the predictor variables, log TP, on their respective means, a step that aids convergence of the algorithm.

Though the means differ, the resultant posterior distributions for $\alpha_{1\text{–}4}$ exhibit considerable overlap (Fig. 5.4) as do the posterior distributions for $\beta_{1\text{–}4}$ (Fig. 5.4), and the respective summary statistics for these parameters confirm these similarities among the group-specific parameters (Table 5.2). Predictive distributions for chlorophyll *a* (using a log TP value of 2.5) in each ecoregion also overlap (Fig. 5.5) with similar summary statistics (Table 5.2).

Additionally, a comparison of the summary statistics for the posterior density of $\sigma$ from the fully pooled model (Table 5.1) and the ecoregion multilevel model (Table 5.2) reveals that they are almost identical. Together, this body of evidence suggests that the ecoregion multilevel

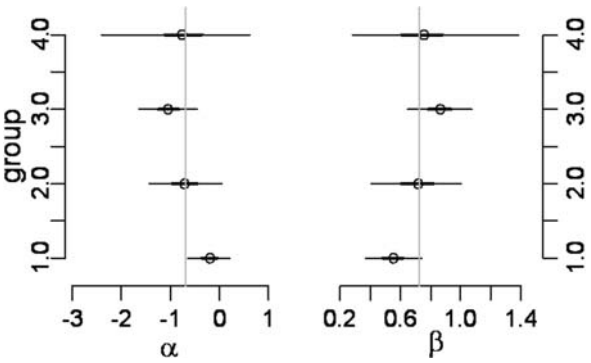



**Fig. 5.4** Parameter estimates for the ecoregion multilevel model. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

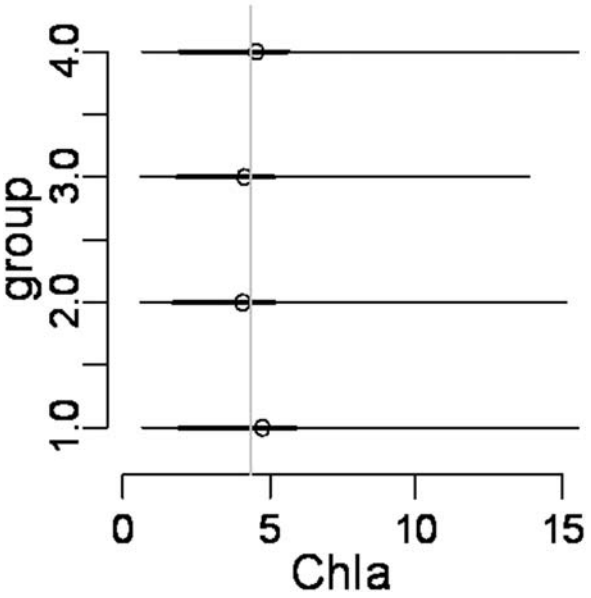



**Fig. 5.5** Chlorophyll *a* predictive distribution for the ecoregion multilevel model. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

model might not represent much improvement over the fully pooled model (Tables 5.1 and 5.2). However, respective DICs of 930.2 and 926.0 for the fully pooled and ecoregion-based models indicate that a slight advantage is conferred by the more complex structure of this particular multilevel model. While this should not be interpreted to mean that the ecoregion-based model is a better forecasting tool, it does imply that the ecoregion model does somewhat better at predicting the existing data set than the fully pooled model does.

Although the ecoregion-based multilevel model does not represent a major improvement over the fully pooled model, it does illustrate some potentially important capabilities of multilevel models. In this model,

**Table 5.2** Summary statistics for ecoregion model – gamma prior

| Quantity | Mean | Median | Sd | 2.5% | 97.5% |
|---|---|---|---|---|---|
| $\alpha_1$ | −0.20 | −0.20 | 0.23 | −0.64 | 0.22 |
| $\alpha_2$ | −0.71 | −0.72 | 0.38 | −1.43 | 0.06 |
| $\alpha_3$ | −1.04 | −1.04 | 0.30 | −1.64 | −0.45 |
| $\alpha_4$ | −0.77 | −0.72 | 0.72 | −2.41 | 0.61 |
| $\beta_1$ | 0.55 | 0.55 | 0.09 | 0.37 | 0.74 |
| $\beta_2$ | 0.72 | 0.72 | 0.16 | 0.41 | 1.00 |
| $\beta_3$ | 0.86 | 0.86 | 0.11 | 0.65 | 1.07 |
| $\beta_4$ | 0.76 | 0.74 | 0.26 | 0.29 | 1.38 |
| $\sigma$ | 0.81 | 0.81 | 0.03 | 0.75 | 0.86 |
| Predicted chlorophyll a 1 | 4.25 | 3.13 | 4.06 | 0.74 | 14.08 |
| Predicted chlorophyll a 2 | 4.02 | 3.04 | 3.57 | 0.55 | 13.32 |
| Predicted chlorophyll a 3 | 4.39 | 3.24 | 4.19 | 0.64 | 14.12 |
| Predicted chlorophyll a 4 | 4.33 | 2.94 | 4.52 | 0.60 | 16.69 |
| DIC | 926.0 | | | | |

ecoregion group 4 (ecoregion 55) contains data from only one lake; therefore the group 4 parameters are based on extremely sparse group-specific data. This is possible because partial pooling in multilevel estimation permits information sharing among the groups. In this case, considerable information is borrowed from the other groups to estimate the group 4 parameters. Thus, multilevel models represent a powerful approach to estimate group specific, or in this example even lake specific models with limited group or lake specific information. However, there are some important caveats associated with this capability, some of which are subtle. Conceptually, it seems apparent that estimating a model with one or only a few observations could be risky, particularly if the observations represent unusual circumstances for that group or lake. And because this is a relatively new approach, it has not been extensively applied and tested to evaluate whether it provides better predictions in this particular context. But in addition, Gelman (2006) warns that some non-informative priors for the variance can actually be quite informative when groups contain sparse data or when there are only a few groups in the multilevel model. In such instances, Gelman (2006) recommends using a uniform prior for $\sigma$ instead of a gamma prior for $1/\sigma^2$ to ensure that the prior is truly non-informative. When this change is made the differences, in this instance, are slight, but worth noting.

With the new non-informative prior, the estimated means and medians for $\alpha_{1-4}$ and $\beta_{1-4}$ differ slightly (Table 5.3) from those resulting from the gamma prior (Table 5.2) while $\sigma$ is essentially unchanged. However, the standard deviation of $\alpha_4$ increases from 0.72 (Table 5.2) to 1.14 (Table 5.3). The overall change in the parameter distributions is visually more apparent by comparing Figs. 5.4 and 5.6. As a result of the increased uncertainty in $\alpha_4$, the width of the 95% credible interval for a prediction

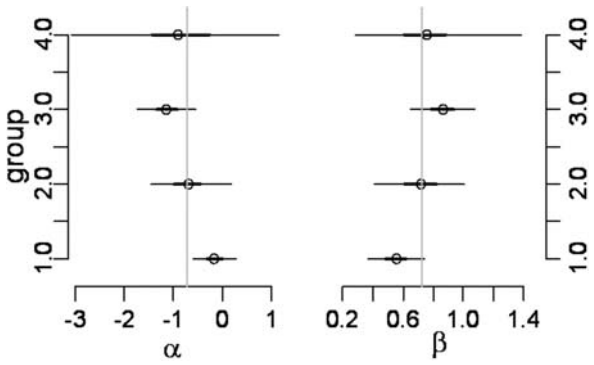



**Fig. 5.6** Parameter estimates for the ecoregion multilevel model, using a uniform prior distribution for sigma. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

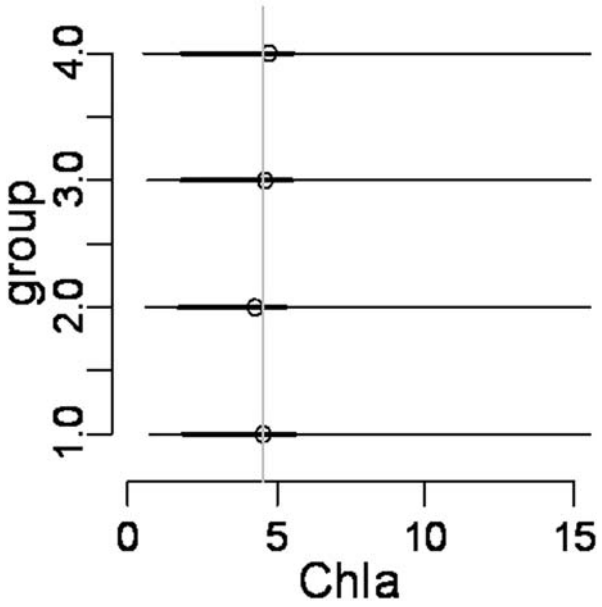



**Fig. 5.7** Chlorophyll *a* predictive distribution for the ecoregion multilevel model, using a uniform prior distribution for sigma. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

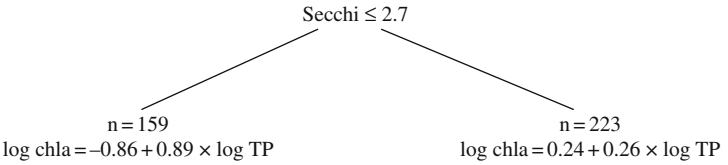
of chlorophyll *a* also increases for group 4, while the prediction for groups 1–3 change little (Table 5.3 and Fig. 5.7). Thus this example illustrates that the gamma prior, in groups with few observations, tends to be somewhat informative and may cause the parameter and predictive uncertainty to be under-estimated in those groups. Additionally, the DIC goes up from 926.0 using the gamma prior (Table 5.2) to 926.5 with the uniform prior (Table 5.3), indicating that the improved performance of the ecoregion-based model relative to the completely pooled model (DIC = 930.2, Table 5.1), may have been slightly overstated.

**Table 5.3** Summary statistics for ecoregion model – uniform prior

| Quantity | Mean | Median | sd | 2.5% | 97.5% |
|---|---|---|---|---|---|
| $\alpha_1$ | −0.17 | −0.16 | 0.21 | −0.56 | 0.23 |
| $\alpha_2$ | −0.70 | −0.7 | 0.41 | −1.51 | 0.15 |
| $\alpha_3$ | −1.12 | −1.14 | 0.30 | −1.70 | −0.50 |
| $\alpha_4$ | −0.85 | −0.80 | 1.14 | −3.1 | 1.1 |
| $\beta_1$ | 0.55 | 0.55 | 0.09 | 0.37 | 0.74 |
| $\beta_2$ | 0.72 | 0.72 | 0.16 | 0.41 | 1.01 |
| $\beta_3$ | 0.86 | 0.86 | 0.11 | 0.65 | 1.07 |
| $\beta_4$ | 0.76 | 0.74 | 0.26 | 0.29 | 1.38 |
| $\sigma$ | 0.81 | 0.81 | 0.03 | 0.75 | 0.87 |
| Predicted chlorophyll a 1 | 4.44 | 3.26 | 3.91 | 0.71 | 15.29 |
| Predicted chlorophyll a 2 | 4.13 | 2.30 | 3.96 | 0.72 | 15.15 |
| Predicted chlorophyll a 3 | 4.27 | 3.20 | 3.73 | 0.60 | 14.09 |
| Predicted chlorophyll a 4 | 4.70 | 3.17 | 5.01 | 0.61 | 19.05 |
| DIC | 926.5 | | | | |

### 5.5.3 *Bayesian Treed Model*

The Bayesian Treed model search resulted in the following tree structure:




This tree structure indicates that the 382 lakes in the data set can be partitioned into two groups, 159 with a secchi depth of $\leq 2.7$ meters and 223 with a secchi depth of $> 2.7$ meters, and that these two groups differ in their log chlorophyll *a*:log TP relationship (Fig. 5.8). This outcome is consistent with results presented by Webster et al. (2008) which showed that regression models predicting chlorophyll *a* from TP are improved with water color (a strong determinant of secchi depth) included as a predictor variable.

Using this result we developed the model:

$$\log(\text{chla}) = \alpha_j + \beta_j \log(\text{TP}) + \varepsilon \tag{5.15}$$

$$\varepsilon \sim \text{N}(0, \sigma^2) \tag{5.15a}$$

where $j = 1$–2, with 159 and 223 lakes in groups 1 (turbid) and 2 (clear), respectively. We built a multilevel model, using these two groups, in a manner analogous to the structure used for the ecoregion-based model. For this example we, again, centered the predictor variables, log TP, on their respective means

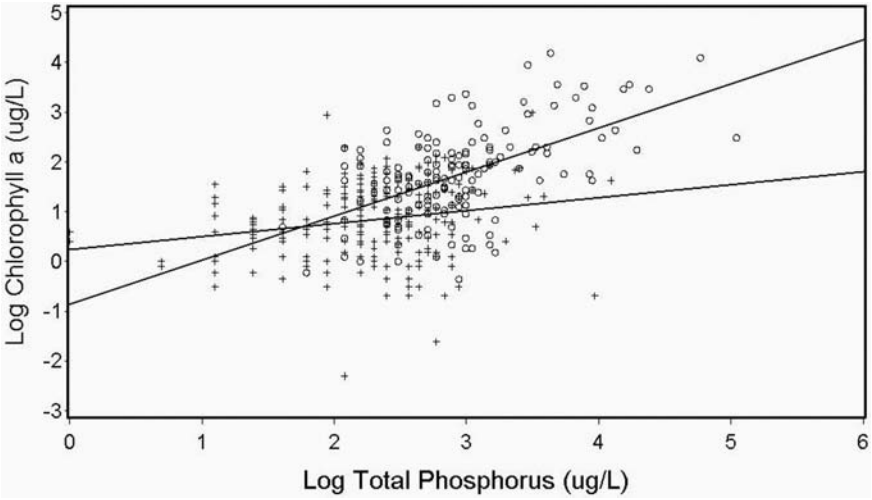



**Fig. 5.8** Log Chlorophyll *a* as a function of log Total Phosphorus for the turbid lakes ($n = 159$, *circles, steeper line*) and clear lakes ($n = 223$, *plus signs, less steep line*)

(note this may cause the parameter estimates to differ slightly from those depicted in the displayed tree) and used a uniform prior for $\sigma$ instead of the gamma prior on $1/\sigma^2$.

In this case, a comparison of the parameter posterior distributions indicates that $\alpha_1$ and $\alpha_2$ have dissimilar means (Table 5.4) and minimal overlap (Fig. 5.9). Similarly, the posterior distributions for $\beta_1$ and $\beta_2$ overlap minimally (Fig. 5.9) which is confirmed by comparing their respective means and standard deviations (Table 5.4). The mean value of $\sigma$ in this model = 0.75 (Table 5.4), a reduction from 0.82 and 0.81 (Tables 5.1 and 5.3) for the fully pooled and ecoregion-based models, suggesting a better fit to the data. This is supported by a DIC of 874.6 (Table 5.4) as compared to DIC values of 930.2 and 926.5 for the first two models. Based on this evidence, the secchi depth-based model

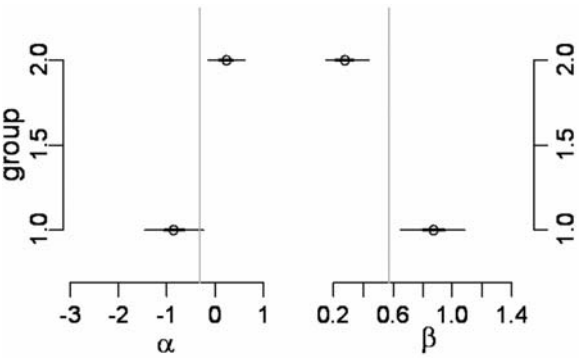



**Fig. 5.9** Parameter estimates for the multilevel model based on secchi depth, identified with the Bayesian Treed Model search. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

**Table 5.4** Summary statistics for secchi model – uniform prior

| Quantity | Mean | Median | sd | 2.5% | 97.5% |
|---|---|---|---|---|---|
| $\alpha_1$ | −0.86 | −0.87 | 0.32 | −1.46 | −0.23 |
| $\alpha_2$ | 0.24 | 0.243 | 0.19 | −0.14 | 0.60 |
| $\beta_1$ | 0.87 | 0.88 | 0.11 | 0.65 | 1.08 |
| $\beta_2$ | 0.27 | 0.27 | 0.081 | 0.12 | 0.44 |
| $\Sigma$ | 0.75 | 0.75 | 0.03 | 0.70 | 0.81 |
| Predicted chlorophyll *a* 1 | 5.36 | 4.13 | 4.55 | 0.93 | 17.51 |
| Predicted chlorophyll *a* 2 | 3.30 | 2.46 | 2.93 | 0.60 | 11.09 |
| DIC | 874.6 | | | | |

suggested by the Bayesian Treed results appears to be the best model, among those tested, for these data.

But is there a practical importance to these secchi depth-based differences, and how would using this model result in decisions that differ from decisions based on the fully pooled model? Suppose, for example, we are charged with evaluating whether or not a log TP concentration of 2.5 (12.2 μg/L in the natural metric) is sufficient to meet a state chlorophyll *a* criterion of 10 μg/L. Guidance provided by the U.S. Environmental Protection Agency suggests that numerical criteria should be regarded as an acceptable 90th percentile, recognizing that ambient concentrations vary spatially and temporally (U.S. Environmental Protection Agency 2000). Thus, to answer this question, we need to evaluate if these models differ in predicting whether the probability of exceeding 10 μg/L chlorophyll *a* is greater than 10%.

Both the mean and median predicted chlorophyll *a* values for secchi groups 1 and 2 are well below 10 μg/L (Table 5.4), as are the mean and median values for the fully pooled model (Table 5.1). Considering only these values of central tendency, which is often done in practice, would result in the conclusion that both models support the notion that 12.2 is an acceptable target TP concentration to meet the chlorophyll *a* criterion. Further, chlorophyll *a* predictive distributions for groups 1 and 2 from the secchi depth-based model exhibit considerable uncertainty, signifying a wide range of possible chlorophyll *a* values. They also display considerable overlap (Fig. 5.10) suggesting minimal differentiation of predicted outcomes. However, the probability of exceeding 10 μg/L for group 1 is ∼12% while for group 2 this probability is ∼3%. The fully pooled model, on the other hand, predicts an 8% probability that chlorophyll *a* will exceed 10 μg/L at a 12.2 μg/L TP level. Thus, according to the fully pooled model 12.2 μg/L is an acceptable target TP concentration for all lakes, while the secchi-based model indicates that only lakes in group 2 will meet that criterion at this TP concentration. Based on the much lower DIC value for the secchi-based model we would likely place more credence in this model's predictions.

Suppose, however, we want to evaluate a lake for which secchi depth measurements are not readily available, so that group membership is unclear. Using

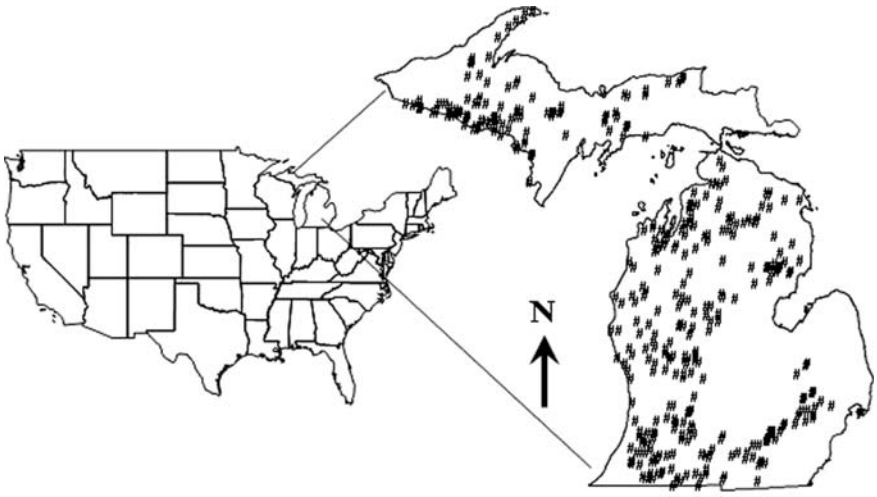



**Fig. 5.10** Chlorophyll *a* predictive distribution for multilevel model based on secchi depth, identified with the Bayesian Treed Model search. The *circle* represents the estimated mean, the *thick black line* is the 50% credible interval and the *thin black line* is the 95% credible interval

the fully pooled model this would not be a problem, because knowing group membership is unnecessary. But the preferred model in this case is the secchi-based model and there are two possible parameter sets available – which of these is more appropriate for a lake with unknown group membership?

In principle, the answer to this question is that we would use the hyper-parameters, which are defined by the prior distributions on $\alpha_j$ and $\beta_j$. In practice, Gelman and Hill (2007) warn that, with only two groups, the hyper-parameters will not be well-estimated and are likely to be overly variable. Thus, the advantage of a two-group multilevel model is limited in this respect. Fortunately, at least for this example, secchi depth is probably the easiest limnological measurement to obtain.

## 5.6 Summary

Our example illustrates some advantages as well as some limitations of using Bayesian multilevel models for inference and prediction with cross-system lake data. Multilevel modeling is a rigorous basis for partial pooling of information among similar, but non-identical groups, and Bayesian approaches provide a framework for uncertainty analysis, an important ingredient for environmental

decision support. The explicit inclusion of quantified uncertainty, expressed probabilistically, provided by Bayesian predictive distributions, underscores the reality that model predictions are probably better regarded as testable hypotheses rather than forecasts of the future.

The importance of models and their limitations were explicitly recognized in the development of the Adaptive Management concept (Holling 1978, Lee 1993). Adaptive Management views environmental stewardship as an ongoing process; models offer guidance for decision-making, and management actions serve as an ecosystem-scale experiment to learn more about system behavior. Measuring the response of the ecosystem to this management experiment supplies data to confirm and update the model and refine future projections. The adaptive process is not well-facilitated by the rather static view of classical null hypothesis-testing in which decisions are limited to accept or reject. However, these ideas are consistent with Bayesian inference, and Bayes theorem provides a rigorous framework for model updating and refinement to implement the Adaptive Management paradigm.

## References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, In Proc. 2nd Int. Syp. Information Theory (eds. B.N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. Chapman & Hall. New York, NY.

Burnham, K. P., and D. R. Anderson. 1998. Model Selection and Inference, New York: Springer.

Canfield, D. E., and R. W. Bachmann. 1981. Prediction of total phosphorus concentrations, chlorophyll *a*, and secchi depths in natural and artificial lakes. Canadian Journal of Fisheries and Aquatic Sciences **38**: 414–423.

Chambers, J. M., A. E. Freeny, and R. M. Heiberger 1992. Analysis of Variance: designed experiments, in /*Statistical Models in S*/, J. M. Chambers and T. J. Hastie (eds.), Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

Chipman, H. A., E. I. George, and R. E. McCulloch. 1998. Bayesian CART model search. Journal of the American Statistical Association **93**: 935–948.

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

Clark, L. A., and D. Pregibon, 1992. Tree Based Models, in *Statistical Models in S*, J. M. Chambers and T. J. Hastie (eds.), Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

Cole, J., G. Lovett, and S. Findlay (eds.). 1991. Comparative analyses of ecosystems: patterns, mechanisms and theories. Springer-Verlag. 375 pp.

Dennis, B. 1996. Discussion: Should ecologists become Bayesians? Ecological Applications **6**: 1095–1103.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters **7**: 509–520.

Freeman, A. M., E. C. Lamon, and C. A. Stow. 2008. Regional nutrient and chlorophyll *a* relationships in lakes and reservoirs: A Bayesian TREED model approach. Ecological Modelling. in press.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis **1**: 515–533.

Gelman, A., and J. Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, NY.

Hession, W. C., D. E. Storm, S. L. Burks, M. D. Smolen, R. Lakshminarayanan, and C. T. Haan. 1995. Using Eutromod with a GIS for establishing total maximum daily loads to Wister Lake, Oklahoma. Pages 215–222 *in* K. Steele, editor. Animal Waste and the Land-Water Interface. Lewis, Boca Raton.

Holling, C. S. ed. 1978. Adaptive Environmental Assessment and Management. John Wiley & Sons. NY.

Hubbard, R., and M. J. Bayarri. 2003. Confusion over measures of evidence ($p$'s) versus errors ($\alpha$'s) in classical statistical testing. American Statistician **57**: 171–178.

Hubbard, R., and J. S. Armstrong. 2006. Why we don't really know what "statistical significance" means: A major educational failure. Journal of Marketing Education **28**: 114–120.

Lamon E. C., O. Malve, and O-P. Pietiläinen. 2008. Lake classification to enhance prediction of eutrophication endpoints in Finnish lakes, Environmental Modeling and Software. **23**: 947–948.

Lamon, E. C., and C. A. Stow. 1999. Sources of variability in microcontaminant data for Lake Michigan salmonids: Statistical models and implications for trend detection. Canadian Journal of Fisheries and Aquatic Sciences **56**, Supplement 1: 71–85.

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

Lathrop, R. C., S. R. Carpenter, C. A. Stow, P. A. Soranno, and J. C. Panuska. 1998. Phosphorus loading reductions needed to control blue-green algal blooms in Lake Mendota. Canadian Journal of Fisheries and Aquatic Sciences **55**: 1169–1178.

Lee, K. N. 1993. Compass and Gyroscope. Island Press. Washington, DC.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing **10**: 325–337.

Magnuson J. J., W. M. Tonn, A. Banerjee, J. Toivonen, O. Sanchez, M. Rask. 1998. Isolation vs. extinction in the assembly of fishes in small northern lakes. Ecology **79**: 2941–2956

Malve, O., and S. S. Qian. 2006. Estimating nutrients and chlorophyll *a* relationships in Finnish lakes. Environmental Science & Technology **40**: 7848–7853.

Martin, S. L., and P. A. Soranno. 2006. Defining lake landscape position: Relationships to hydrologic connectivity and landscape features. Limnology and Oceanography **51**: 801–814.

National Research Council. 2001. Assessing the TMDL Approach to Water Quality Management. National Academy Press, Washington, D.C.

Pappenberger, F., and K. J. Beven. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resources Research **42**. Article number W05302.

Qian, S. S., C. A. Stow, and M. E. Borsuk. 2003. On Monte Carlo methods for Bayesian inference. Ecological Modelling **159**: 269–277.

Qian, S. S., and C. W. Anderson. 1999. Exploring factors controlling the variability of pesticide concentrations in the Willamette River Basin using tree-based models. Environmental Science & Technology **33**: 3332–3340.

Reckhow, K. H. 1988. Empirical-models for trophic state in Southeastern United-States lakes and reservoirs. Water Resources Bulletin **24**: 723–734.

Reckhow, K. H. 1990. Bayesian-inference in non-replicated ecological studies. Ecology **71**: 2053–2059.

Reckhow, K. H. 1993. Random coefficient model for chlorophyll nutrient relationships in lakes. Ecological Modelling **70**: 35–50.

Reckhow, K. H. 1996. Improved estimation of ecological effects using an empirical Bayes method. Water Resources Bulletin **32**: 929–935.

Reckhow, K. H., and S. C. Chapra. 1983. Engineering approaches for lake management, Volume 1: Data Analysis and Empirical Modeling. Butterworth Publishers. Boston.
Rohm C. M., J. M. Omernik, A. J. Woods, and J. L. Stoddard. 2002. Regional characteristics of nutrient concentrations in streams and their application to nutrient criteria development. Journal of the American Water Resources Association **38**: 213–239.
Salsburg, D. 2001. The Lady Tasting Tea. Henry Holt and Company. NY.
Schwarz, G., 1978. Estimating the dimension of a model, Annals of Statistics **6**:461–464.
Sonquist, J. N., and J. N. Morgan. 1964. The Detection of Interaction Effects. Monograph 35, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B **64**: 583–639.
Stow, C. A., S. R. Carpenter, and R. C. Lathrop. 1997. A Bayesian observation error model to predict cyanobacterial biovolume from spring total phosphorus in Lake Mendota, Wisconsin. Canadian Journal of Fisheries and Aquatic Sciences **54**: 464–473.
Stow, C. A., S. R. Carpenter, K. E. Webster, and T. M. Frost. 1998. Long-term environmental monitoring: Some perspectives from lakes. Ecological Applications **8**: 269–276.
Stow, C. A., K. H. Reckhow, and S. S. Qian. 2006. A Bayesian approach to retransformation bias in transformed regression. Ecology **87**: 1472–1477.
Stow, C. A., and D. Scavia. 2008. Modeling Hypoxia in the Chesapeake Bay: Ensemble estimation using a Bayesian hierarchical model. Journal of Marine Systems. In press.
U.S. Environmental Protection Agency. 2000. Nutrient Criteria Technical Guidance Manual, Lakes and Reservoirs. Office of water. EPA 822-B00-001.
Vollenweider, R. A. 1968. The Scientific Basis of Lake and Stream Eutrophication with Particular Reference to Phosphorus and Nitrogen as Eutrophication Factors. Technical Report DAS/DSI/68.27. OECD, Paris.
Vollenweider, R. A. 1969. Possibilities and limits of elementary models concerning the budget of substances in lakes. Archiv für Hydrobiologie **66**: 1–36.
Vollenweider, R. A. 1975. Input-output models with special reference to the phosphorus loading concept in limnology. Schweizerische Zeitschrift fur Hydrologie **37**: 53–84.
Vollenweider, R. A. 1976. Advances in defining critical loading levels for phosphorus in lakes eutrophication. Mem. Ist. Ital. Idrobiol. **33**:53–83.
Webster, K. E., P. A. Soranno, K. S. Cheruvelil, M. T. Bremigan, J. A. Downing, P. Vaux, T. Asplund, L. C. Bacon, and J. Connor. 2008. An empirical evaluation of the nutrient color paradigm for lakes. Limnology and Oceanography. **53**:1137–1148.
Winkler, R. L., 2003. An Introduction to Bayesian Inference and Decision, Second Edition, Probabilistic Publishing, Gainesville, FL, USA.