Transactions of the American Fisheries Society 137:841–851, 2008
© Copyright by the American Fisheries Society 2008
DOI: 10.1577/T06-142.1

[Article]

Model-Based Clustering in a Brook Trout Classification Study within the Eastern United States

HUIZI ZHANG
Department of Statistics, Virginia Tech University, Blacksburg, Virginia 24061, USA

TERESA THIELING
U.S. Forest Service, Fish and Aquatic Ecology Unit, James Madison University, Mail Stop Code 7801, Harrisonburg, Virginia 22807, USA

SAMANTHA C. BATES PRINS
Department of Mathematics and Statistics, James Madison University, Mail Stop Code 1911, Harrisonburg, Virginia 22807, USA

ERIC P. SMITH*
Department of Statistics, Virginia Tech University, Blacksburg, Virginia 24061, USA

MARK HUDY
U.S. Forest Service, Fish and Aquatic Ecology Unit, James Madison University, Mail Stop Code 7801, Harrisonburg, Virginia 22807, USA

Abstract.—Data collected on the population status (extirpated or present) of brook trout *Salvelinus fontinalis* at the landscape level across the eastern United States is useful for identifying important stressors and predicting brook trout status at the watershed level. However, when dealing with data compiled over a large region, a single model may not adequately describe relationships between variables. To find models with better classification performance, we used a Monte Carlo model-based clustering method with logistic regression models to obtain subregions with good predictive performance. To subdivide the eastern United States into subregions, we used Voronoi tessellations with randomly selected centers. The average fraction correctly classified for fit was the criterion used when searching for optimal models within clusters. Logistic regression models were chosen by stepwise selection based on five explanatory variables: elevation, percentage of forested land, percentage of agricultural land, road density, and an environmental factor (depositional NO_3 and SO_4). Application of the method to the brook trout status data set resulted in six subregions and improved predictive ability by approximately 20% relative to the single regionwide model. Resulting models were also more interpretable than the single model and reflected effects at a smaller spatial scale. In contrast to results of the single model, the role of elevation in the six subregional models was consistent with expectations and indicated an increased probability of brook trout presence with increased elevation. The resulting models should be useful for identification and prioritizing sites for restoration and recovery programs.

In the 1600s, brook trout *Salvelinus fontinalis* in the eastern United States were prevalent from Georgia to Maine. Human perturbations led to the extirpation of brook trout from many of the streams in which they existed (MacCrimmon and Cambell 1969; Galbreath et al. 2001). According to Hudy et al. (2006), anthropogenic physical, chemical, and biological perturbations have resulted in a significant loss (>50%) of self-sustaining brook trout populations within 59% of eastern U.S. subwatersheds, raising concerns among numerous state and federal agencies, nongovernment organizations, and anglers.

Many of the brook trout extirpations occurred in the early 1900s due to logging and agricultural practices. Changes over the last 100 years were often drastic and included construction of over 75,000 dams (USACE 1998) and 2 million miles of roads (Navtech 2001) and a human population increase of 90 million residents (U.S. Census Bureau 2002). The result has been dramatic land use changes; currently, over 30% of the average subwatersheds are classified as areas of human land use (USGS 2004).

Understanding the relationships between brook trout population status and perturbations is essential for

* Corresponding author: epsmith@vt.edu

Received June 19, 2006; accepted October 27, 2007
Published online May 1, 2008

developing useful managerial strategies for watershed-level restoration, inventory, and monitoring. To help manage brook trout populations and to prevent further extirpations, it is necessary to investigate possible causative factors and to model their influences on population loss. Prediction of the species' population status at a site is essential for risk management and restoration. Several approaches, including laboratory and watershed studies and large-scale assessment, can be used to study factors affecting population loss. Small-scale assessments are likely to be useful for individual sites but may not be applicable over the entire range of the species (see Rashleigh et al. 2005). Large-scale assessments for many aquatic species have been useful in identifying and quantifying problems, information gaps, restoration priorities, and funding needs (Williams et al. 1993; Davis and Simon 1995; Frissell and Bayles 1996; Warren et al. 1997; Master et al. 1998). Previous projects at the landscape scale have examined bull trout *Salvelinus confluentus* (Rieman et al. 1997) and Pacific salmon *Oncorhynchus* spp. (Thurow et al. 1997).

The importance of landscape-level brook trout analysis has been discussed (e.g., Rieman et al. 1997; Kocovsky and Carline 2006) at the state and basin levels. Using a landscape-level subwatershed analysis for the eastern United States, Hudy et al. (2006) and Thieling (2006) developed a brook trout data set and Thieling (2006) investigated a variety of modeling approaches for predicting extirpation of brook trout. Using classification trees, Thieling (2006) was able to develop a model that produced reasonably good prediction with four to five potentially causative variables. While correct classification rates were adequate for these models, several regions had relatively weak classification rates, namely sites in eastern Pennsylvania and some sections of North Carolina, Georgia, and Tennessee. Lower classification rates in some regions suggest that different models apply in these regions.

We believe that an important component of study design is the selection of scale for the data analysis. With large-scale studies, it seems reasonable that models developed over regional subsets would more accurately describe brook trout population relationships than a single model applied over a large area. When the spatial extent is quite large, it is reasonable to expect that the model will differ among different regions. For example, the effect of agricultural practices may differ between northern and southern regions, resulting in different stress models. Effects of some stressors may vary from higher to lower elevations. One approach used to account for location differences involves dividing the entire data set into groups or clusters (e.g., ecoregions, states), such that the observed measurements of interest are similar within each cluster and dissimilar among clusters, and then applying separate models accordingly. This approach is the basis of biological monitoring procedures such as the River Invertebrate Predication and Classification System (Wright et al. 1984) and the Australian Rivers Assessment System (Nichols et al. 2000). Users first cluster sites based on natural factors that influence biological conditions and then model the relationships with stressors. Such clustering methods are inadequate for our purposes, since there is no guarantee that the resulting models will correctly identify and classify relationships between presence–absence and relevant management variables. Making groups of sites as different as possible based on natural factors can weaken stressor relationships if the stressor gradient is associated with the natural variables or if the stressor gradient crosses the cluster boundaries. It seems more logical to establish a regional clustering procedure with strong relationships between extirpation, stressors, and physical variables.

We propose a clustering method that uses Voronoi tessellations for dividing the spatial region into clusters and logistic regression for modeling brook trout status within each cluster as a way of finding good models for predicting status. We evaluate the predictive ability of cluster-specific logistic regression models for brook trout status and compare it with the predictive ability of a single model. We also compare the differences and similarities in estimated model parameters.

Methods

Brook trout data.—Hudy et al. (2006) and Thieling (2006) discussed the distribution, status, and threats to brook trout within the eastern United States. The study area covered 16 states stretching from Maine to Georgia, and complete data were available for 3,337 subwatersheds. The candidate stressor metrics included 63 anthropogenic and landscape variables. The response variable we used was self-sustaining brook trout population status (extirpated or present). Sites of extirpation are those subwatersheds from which historically self-sustaining populations of brook trout have been lost. The pattern of extirpation varied considerably over the study region (Table 1). Three states in the study had a nearly uniform response (i.e., brook trout were present in 98% of the locations). With these states excluded, the sample size was 2,789 subwatersheds; brook trout populations were categorized as present in 1,717 of the sites and as extirpated from the remainder.

Statistical methods.—Thieling (2006) screened 63 candidate landscape-level metrics based on redundancy

(i.e., only one variable was retained for further screening from a pair of variables with a correlation greater than 0.80) and significance to the response variable (using Wald's chi-square test from logistic regression and a classification tree search for significant variables). Based on the screened variables, we chose five predictor variables to use for model building: (1) mean elevation (m) of the subwatershed, (2) subwatershed road density (km of road per km^2 of land), (3) percentage of subwatershed area in agricultural use, (4) an environmental factor that incorporated depositional NO_3 and SO_4 (depositional chemistry), and (5) percentage of forested lands in the water corridor of the subwatershed. The Navtech (2001) enhanced Topological Integrated Geographic Encoding and Referencing system data were the basis of the road density calculations. Land use variables were obtained from the 1992 National Land Cover Dataset that was completed for the study area in 1998 (USGS 2004). To measure stress from depositional chemistry, estimates of NO_3 and SO_4 (kg/ha) were obtained by interpolating existing maps (National Atmospheric Deposition Program 2005). We first standardized the NO_3 and SO_4 variables and then summed them to estimate a deposition factor.

As part of the variable processing, we applied the Box–Cox transformation approach and transformed variables as follows: road density (log transformation), agricultural percentage (square-root transformation), depositional chemistry (log transformation), and forest percentage (square-root transformation). We further centered and scaled all five variables using the mean and standard deviation calculated from the reference (presence) group to allow better interpretation of variable importance.

Our clustering method was a variation on the method of S. C. B. Prins (unpublished) and had three steps. First, we partitioned the space using a randomly selected set of points from within the region. Second, we fitted a stepwise logistic model within each partition. Finally, we evaluated the quality of the fit aggregated across the clusters. These three steps were repeated a large number of times, varying the number of partitions, and the best value of the evaluation criterion was used to determine the final regions and models. We iterated over these steps to further improve the fit.

The first step involved partitioning the region into k clusters using Voronoi tessellations and Delaunay triangulation (Møller 1994; Okabe et al. 2000). To do this, we randomly generated k points (labeled $g_1, \ldots, g_k$) corresponding to latitudes and longitudes within the region by selecting an existing site from the data set and then adding a random value from the uniform distribution, $U(-\tau, \tau)$, where $\tau > 0$ is the maximum perturbation considered for the point. In the initial run, τ was set to 0.50. Each simulation thus generated a spatial center for each of the k clusters. We then identified the group of sites that were closest to each of these centers. In the context of spatial data, the distance between sites i and j can be measured in several ways, depending on how curvature of the Earth is accounted for (Banerjee 2004). Euclidian distance assumes that the points are on a flat surface. We used geodetic distance (Banerjee 2004), which assumes that the Earth is a sphere. The formula is

$$D(\mathbf{L}_i, \mathbf{L}_j) = R \arccos[\sin L_{i2} \sin L_{j2} + \cos L_{i2} \cos L_{j2} \cos(L_{j1} - L_{i1})],$$

where $\mathbf{L}_i = (L_{i1}, L_{i2})$ represents the vector of longitude (L_{i1}, converted to radians) and latitude (L_{i2}) for the centroid of site i; $\mathbf{L}_j$ is similarly defined for the centroid of site j, and R is the radius of the Earth (approximately 6,371 km). We calculate $D(\mathbf{L}_i, \mathbf{g}_j)$ for all sites with locations $\mathbf{L}_i$ ($i = 1, 2, \ldots n$) and cluster centers $\mathbf{g}_j$ ($j = 1, 2, \ldots k$) and assign a site to the closest cluster.

Figure 1 illustrates the tessellation process for a single simulation. The open circles in the figure represent the seed points or spatial centers. The

TABLE 1.—Summary of the locations of 3,337 subwatersheds for which self-sustaining population status (extirpated or present) of brook trout was examined by model-based clustering (N = number of subwatersheds sampled in each state). Three states (New Hampshire, Vermont, and Maine) had a nearly uniform response (i.e., brook trout presence in 98% of subwatersheds) and were therefore not modeled; the remaining states were modeled via logistic regression.

State	N	Status	
		Extirpated	Present
Unmodeled states			
New Hampshire	47	0	47
Vermont	186	6	180
Maine	315	5	310
Total	548	11	537
Modeled states			
Connecticut	175	29	146
Massachusetts	130	20	110
New York	350	115	235
Pennsylvania	1,085	444	641
New Jersey	58	31	27
Ohio	4	1	3
Maryland	132	82	50
Virginia	319	148	171
West Virginia	174	24	150
South Carolina	19	12	7
North Carolina	214	95	119
Tennessee	54	18	36
Georgia	75	53	22
Total	2,789	1,072	1,717

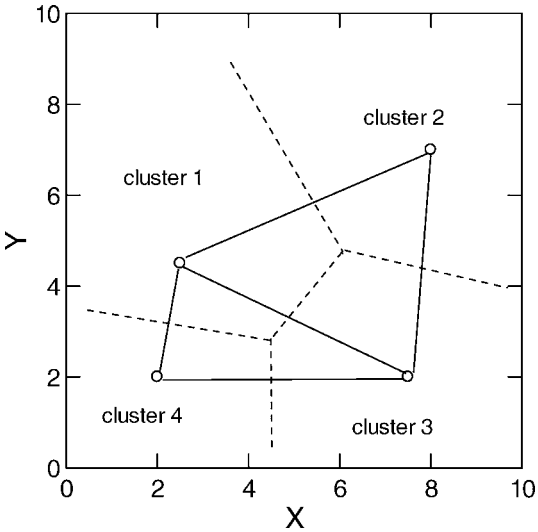

FIGURE 1.—Illustration of the Voronoi tessellation method for a spatial region corresponding to a collection of brook trout subwatersheds with coordinates given by X (longitude) and Y (latitude). Open circles represent seed points or spatial centers that were randomly generated within the region. Triangles (solid lines) represent the Delaunay triangulation of the region based on the generated seed points. Boundaries (dashed lines) for the different clusters or subregions bisect the sides of the triangles and define the Voronoi tessellations. Subwatersheds with coordinates inside the boundaries are assigned to that cluster.

triangles form the Delaunay triangulation of the region. The boundaries (dashed lines) for the different clusters or subregions bisect the sides of the triangles and define the Voronoi tessellations. Points within a given cluster are closer to the center of this cluster than to the center of any other cluster.

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

$$\begin{aligned}\text{logit}(p) &= \log\left(\frac{p}{1-p}\right) = x\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m,\end{aligned}$$

where x_1 through x_m are the explanatory variables and $\boldsymbol{\beta}$ is a vector of $m + 1$ regression parameters (β_0 corresponds to the intercept term). Logistic regression models are useful for management purposes, because parameter estimates measure the importance of a variable relative to the other variables and can assist in selecting appropriate management strategies. Statistically, the model has better asymptotic efficiency than other nonparametric or semiparametric approaches (Pepe et al. 2005) and is quite robust against link violation (Li and Duan 1989). We used the m predictor variables to fit the cluster-specific stepwise logistic regression models to data in each of the k clusters. A stepwise procedure was used to eliminate redundant variables.

The third step involved choosing a performance measure for the clustering solution. Many measures are possible. Because our focus was on predictive models, the measure we used was the average fraction correctly classified for fit (AFCCF; Wilkinson 1999). The AFCCF is a measure of the classification model's ability to make a correct prediction; it estimates the proportion of correctly classified sites without using a cutoff for prediction. The formula for AFCCF is

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

where y_i is the value of the ith observation ($y_i = 0$ for brook trout presence; $y_i = 1$ for extirpation), $\hat{p}_i$ is the estimated probability of extirpation, and n is the number of sites. Note that the AFCCF will be between 0 and 1 inclusive. A value of $\hat{p}_i$ that is high for sites of extirpation and low for sites of brook trout presence indicates a model with good classification. In this case, the AFCCF will have a value closer to 1.0. The AFCCF is similar to the coefficient of determination (R^2) for regression models in that higher values indicate a stronger relationship between the model and the data. We used 10-fold cross validation to correct for possible bias from using the same data set to test model accuracy and to fit the model (Hastie et al. 2001). We used AFCCF rather than a model-building criterion, such as Akaike's information criterion, since the focus was on prediction rather than model building.

The three-step process was repeated many times (number of simulations $S = 10{,}000$) in an attempt to find near-optimal partitioning. To further improve the fit, we iteratively searched neighborhoods around the near-optimal value. Random numbers generated from a uniform distribution with a smaller parameter ($\tau = 0.25$ or 0.10) were added to the cluster centers from the initial optimal fit. We used these perturbed centers as seeds and repeated the process 100 times. We also used the same procedure on other near-optimal results to search for better models.

A potential problem with fitting a logistic regression model to partitioned data is that complete separation between status categories or a lack of convergence is possible. With logistic regression models, these problems are often due to small sample sizes or regions that contain one type of status (e.g., extirpation or

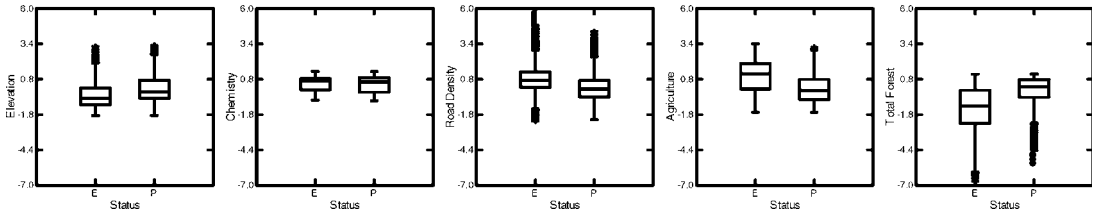

FIGURE 2.—Box plots of five transformed predictor variables (elevation [m]; depositional NO_3 and SO_4 [chemistry]; road density [km road/km^2 land]; percent area in agricultural use [agriculture]; and percentage of forested lands in the water corridor [total forest]) used in model-based clustering of brook trout population status (extirpated [E] or present [P]) in subwatersheds of the eastern USA. The box indicates the 75th and 25th percentiles; the middle bar represents the median; whiskers extend to the largest (and smallest) observations within 1.5 times the interquartile range of the box where interquartile range is the range of the 75th and 25th percentiles. Extreme values are indicated by asterisks.

presence) almost exclusively (i.e., the response is essentially uniform over the subregion). To reduce the occurrence of complete separation, we used a minimum sample size of 100 for the logistic regressions. Smaller minimum sample sizes tended to result in at least one cluster containing subwatersheds with mostly uniform status. Models fit to these data may result in a high AFCCF for the cluster but do not necessarily have a good explanatory ability.

We repeated the three steps for a range of cluster sizes from 2 to a maximum of K clusters during one randomization–simulation run. The process was repeated S times such that the total number of runs was equal to $K \times S$. After obtaining a solution, we varied the size of S to ensure closeness to optimality and to evaluate the method's sensitivity to S.

Users of cluster analysis base their selection of the number of groups on both subjective and objective criteria. We selected the optimal partitioning based on the AFCCF criterion, sample sizes (extirpated and present) for each group, and interpretation of the results.

Results

Box plots of the transformed variables are displayed in Figure 2 and indicate a large amount of variation and a slight amount of separation. Figure 3 displays the optimal values of the AFCCF criterion for 2–9 clusters and an S-value of 10,000. The shape of the AFCCF curve increases quickly, then levels at around six clusters. Cross validation did not change the AFCCF values by an appreciable amount.

The six-cluster solution that resulted in the maximum AFCCF was chosen as the final clustering solution. We based our decision on interpretation and the high value of AFCCF, both overall and for individual clusters. Models with additional clusters tended to have a slightly better AFCCF due to the splitting of clusters containing high proportions of presence, but such models did not produce more interpretable models. A further check of the six-cluster solution using 20,000 iterations with additional searches resulted in essentially the same model.

Figure 4 displays the geographical locations of the resulting six clusters. Cluster 1 is a region associated with the northeastern part of New York, Connecticut, Rhode Island, and Massachusetts. This cluster contains the least-disturbed sites and the highest proportion of sites containing brook trout. Cluster 2 consists mostly of sites from western New York State. Agriculture dominates the area, and exotic fishes (e.g., brown trout *Salmo trutta*) may affect the presence of brook trout. Cluster 3 covers eastern Pennsylvania and southeastern

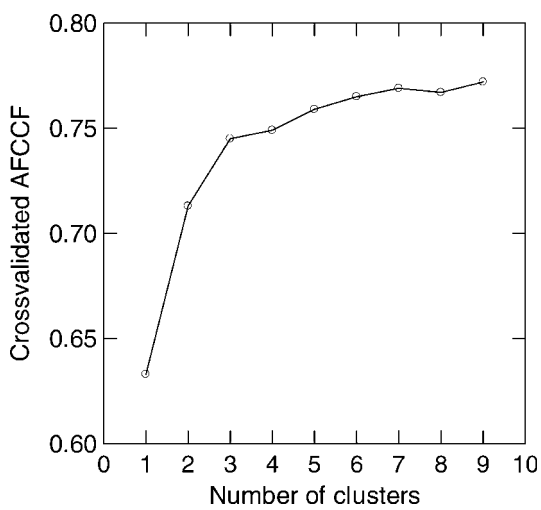

FIGURE 3.—Tenfold cross-validated values of the average fraction correctly classified for fit (cross-validated AFCCF, y-axis) presented in relation to the number of clusters (2–9, x-axis) evaluated in 10,000 simulations; here, AFCCF describes the ability of each set of models to predict brook trout population status (extirpated or present) in eastern U.S. subwatersheds and indicates that six is the optimal number of clusters for use in modeling.

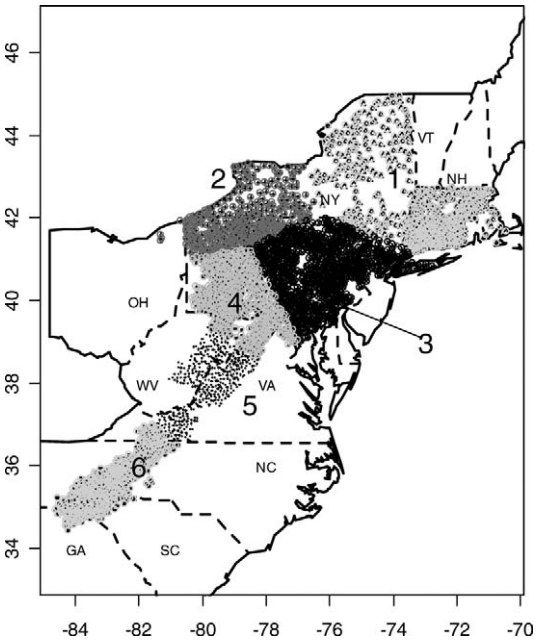

FIGURE 4.—Delineation of six eastern U.S. subregions identified by a model-based clustering approach as providing the best ability to predict brook trout status (extirpated or present) in subwatersheds (shown as points within each subregion; cluster 1 = northeastern part of New York, Connecticut, Rhode Island, and Massachusetts; cluster 2 = western New York State; cluster 3 = eastern Pennsylvania and southeastern New York; cluster 4 = western Pennsylvania and northern parts of Virginia and West Virginia; cluster 5 = northern portion of the southern Appalachian Mountains region; cluster 6 = southern portion of the southern Appalachian Mountains region). The *x*-axis is longitude and the *y*-axis is latitude.

New York. Loss of natural landscapes due to urbanization and loss of forest are probably important factors in this region. Western Pennsylvania and the northern parts of Virginia and West Virginia make up cluster 4. Local geological patterns and mining are likely to be determinants of status in this region. Clusters 5 and 6 correspond to the southern Appalachian Mountains region, divided into north and south, respectively.

Table 2 lists the values of the AFCCF criterion for each cluster. The overall AFCCF was 0.76. The AFCCF values based on 10-fold cross validation results were essentially the same as the regular estimates. Relative to the single-cluster model, the improvement was over 20%, as measured by increase in AFCCF. The AFCCF values tended to be good for all clusters except perhaps cluster 6. Thieling (2006) indirectly identified cluster 6 as an area with lower classification rates than other regions.

TABLE 2.—Average fraction correctly classified for fit (AFCCF) calculations, measuring the predictive ability of six subregional models of brook trout population status (extirpated or present) and a single regionwide model for the eastern USA (observations = subwatersheds with status data). The AFCCF values based on 10-fold cross validation ($AFCCF_{cv}$) are also shown.

Model cluster	Observations	AFCCF	$AFCCF_{cv}$
1	493	0.87	0.86
2	365	0.80	0.80
3	711	0.71	0.71
4	458	0.75	0.74
5	368	0.77	0.77
6	394	0.69	0.68
Overall	2,789	0.76	0.76
Single	2,789	0.63	0.63

Table 3 summarizes the parameter estimates and their significance in the logistic regression models for the model-based clustering and single-model approaches. There was heterogeneity in both the magnitude and sign of the coefficient estimates between the single regionwide model and the six subregional models and among subregional models. Based on the single model, the parameter estimate for elevation was relatively small and positive at 0.211. Elevation is partly an indicator of the water temperature within the subwatershed and is confounded with human development (less in higher elevations) and location of the site (northern sites tend to have lower elevations). The estimated coefficient was positive (0.211), which suggests that higher-elevation streams (which tend to be cooler) are more likely to be sites of extirpation (in our models, positive coefficients indicate increasing probability of extirpation as the value of the explanatory variable increases). The positive value of the coefficient seems to contradict the fact that brook trout prefer higher-elevation, cooler streams, but the estimate is valid because it reflects the larger spatial scale. Brook trout tended to be present in the northern region, where elevation is lower, rather than the southern region, where elevation is higher (see the box plots of elevation in Figure 5). Therefore, the estimate reflected a general reduction in elevation from south to north with corresponding declines in temperature and in the proportion of sites where brook trout are extirpated. Elevation parameter estimates obtained from the model-based clustering approach were uniformly negative (i.e., probability of extirpation decreased as elevation increased) and reflected influences at a smaller spatial extent. This result is consistent with prior local-scale expectations (Hudy et al. 2006).

There were considerable differences among intercept estimates. The exponential of the estimated intercept

TABLE 3.—Parameter estimates for five predictor variables (elevation [m]; road density [km road/km^2 land]; percent area in agricultural use; depositional NO_3 and SO_4; percentage of forested lands in the water corridor) used in a single regionwide logistic model and six subregional models identified from cluster-based modeling of brook trout population status (extirpated or present) in subwatersheds of the eastern USA (number of sites of extirpation [EX] and presence [PR] are shown). Empty cells indicate variables that were not added to the given model. Significance was less than 0.01 except for items marked by asterisks (*$P < 0.05$).

Model	Intercept	Elevation	Road density	Percent agriculture	Percent forest	Depositional chemistry	EX	PR
Single	−1.169	0.211	0.263	0.486	−0.329		1,717	1,072
Cluster 1	−5.469	−1.811	0.954			0.758*	48	445
2	−0.285	−3.642		1.441		−3.620	166	199
3	−2.583	−1.221	0.313*	0.469	−0.438		304	407
4	0.123	−3.581	0.840			−0.981*	255	203
5	−0.404	−1.581	−0.690*		−1.237	−2.919	106	262
6	3.444	−2.272	0.554			5.606	193	201

represents the probability of extirpation at sites with zeros for all variables (i.e., baseline models). Baseline (intercept-only) models for most clusters (clusters 1–3 and 5) favored brook trout presence, whereas cluster 6 had a large positive intercept of 3.444. Cluster 6 represented the southern Appalachian area, and the associated model tended to incorrectly classify sites of extirpation as being sites of brook trout presence. A closer look at the original data in that region revealed some interesting features. Figure 5 presents comparisons of the predictor variables across clusters and status. Cluster 6 represented a subregion of higher elevation (mountain areas), lower proportions of agricultural use, and a higher forested percentage

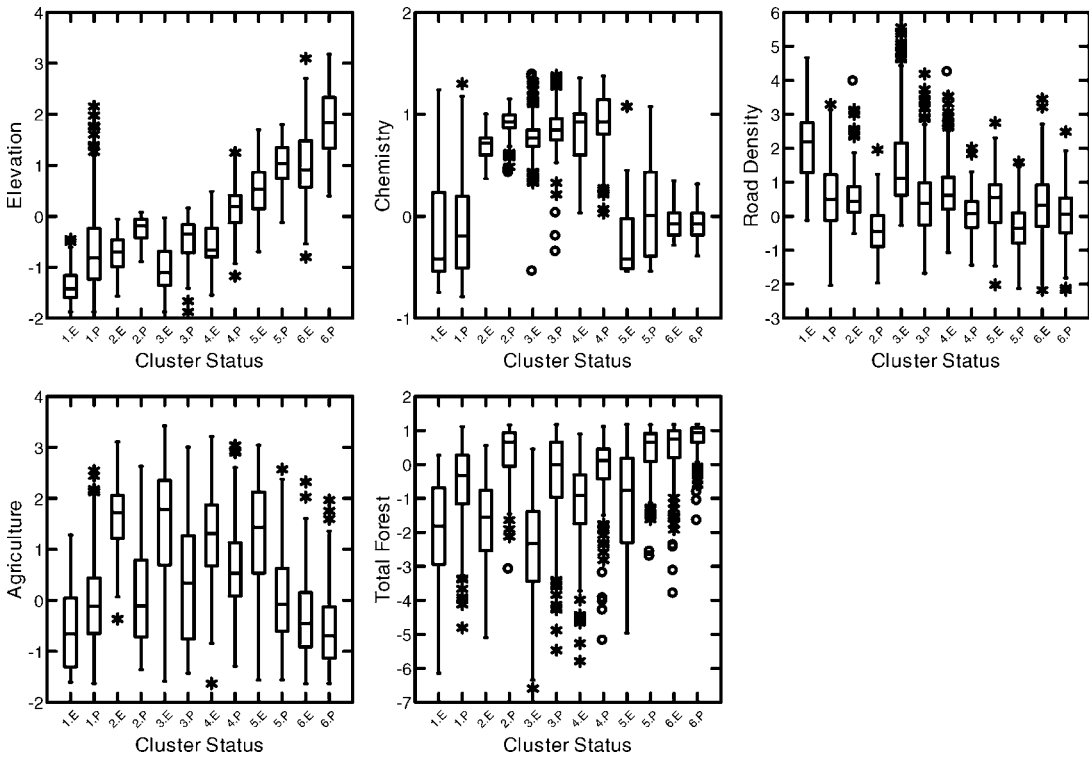

FIGURE 5.—Box plots of five transformed, centered, and scaled predictor variables (elevation [m]; depositional NO_3 and SO_4 [chemistry]; road density [km road/km^2 land]; percent area in agricultural use [agriculture]; and percentage of forested lands in the water corridor [total forest]) used in six subregional models of brook trout population status (extirpated [E] or present [P]) in subwatersheds of the eastern USA. Subregion locations (clusters 1–6) are defined in Figure 4. The *y*-axis defines values of the indicated variable, while the *x*-axis is the combination of cluster number and status. Characteristics of the box plots are described in Figure 1.

relative to the other subregions. Our prior expectation was that this area would favor brook trout presence; hence, we expected the baseline model to have a small or negative intercept. In fact, sites of presence were not predominant in this subregion. Cluster 6 contained 193 sites of extirpation and 201 sites of presence, which suggests that the intercept would be near zero. Instead, our cluster 6 model produced the positive intercept estimate of 3.44, which implies that the general pattern in the area is different from patterns in the other areas and is suggestive of high baseline extirpation. The discrepancy indicates that further investigation is needed to achieve better classification performance. As indicated by Thieling (2006), past land use practice and the stocking of exotic rainbow trout *O. mykiss* into restored subwatersheds have displaced brook trout; therefore, rainbow trout presence is a very important metric affecting brook trout status. Although an invasive species metric was developed, it was not sufficiently accurate for good prediction. By using model-based clustering, we were able to discern such irregular subregions within the larger region and to better control the misclassification rate relative to a single-model approach. In future studies, inclusion of a better exotic fish metric may further enhance the model's predictive performance.

According to Thieling's (2006) retrospective study, the majority of subwatersheds for which extirpation was predicted but that in fact contained brook trout were located in eastern Pennsylvania (cluster 2) and western New York (cluster 3). Thieling's (2006) study predicted extirpation in many watersheds because of high road density and increasing urbanization. However, in practice, many of these watersheds had high-elevation refuges where brook trout were present in small, isolated coldwater habitats.

After clustering, the importance of elevation, agriculture, and forest outweighed that of road density, especially in clusters 2 and 3, and we were able to obtain better prediction for these areas. The depositional chemistry variable seemed especially important in clusters 2 (western Pennsylvania) and 5 (Virginia and West Virginia mountains). Thus, the models generally conformed to expectations associated with knowledge about stressors in the subregions defined by these clusters. Exceptions were the importance of chemistry in clusters 1 and 6, which had rather low levels of deposition. For cluster 1, chemistry became important after we entered elevation into the model; graphical analysis suggested that chemistry entered the model because sites of extirpation were lacking at higher elevations. For cluster 6, graphical analysis indicated that after adjusting for elevation, the probability of extirpation increased with decreasing depositional chemistry. We suspect that in this region, deposition is probably not a causative factor but is associated with the west–east spatial gradient.

Discussion

In this paper, we have described a method for using model-based clustering with spatial data. In particular, we developed algorithms for segmenting binary response data collected over a large region based on the performance of logistic regression models. We used Voronoi tessellation techniques and a predictive criterion as the performance measure with a Monte Carlo search for the optimal clustering solution. Application of this method to a brook trout data set demonstrated its potential for achieving better classification performance than a similar model that ignored clustering.

The importance of this work rests not so much in the models and their included variables but rather in the improvement in predictive ability. Improved prediction will be useful for prioritizing sites according to restoration or preservation. For example, a site where brook trout are present but have a high probability of extirpation is probably worth preserving. A site of extirpation that is predicted to have a low probability of extirpation could be considered a good candidate for brook trout restoration. An improved predictive ability provides a better list of sites for restoration and recovery and hence increases the potential for a more successful management program.

An important caveat is that these models do not incorporate historical information. Many physical, chemical, and biological changes at the watershed level have occurred over the last 200 years in the brook trout's native range in the eastern United States (MacCrimmon and Campbell 1969; Jenkins and Burkhead 1993; Marschall and Crowder 1996; Yarnell 1998). Historic and current land use practices (King 1937, 1939; Lennon 1967; Kelly et al. 1980; Nislow and Lowe 2003); increased water temperature (Meisner 1990); the spread of exotic and nonnative fishes (Moore et al. 1983; Larson and Moore 1985; Moore and Ridley 1986; Strange and Habera 1998); fragmentation of habitats by dams and roads (Belford and Gould 1989; Gibson et al. 2005); changes in water quality (Fiss and Carline 1993; Gagen et al. 1993; Clayton et al. 1998; Hudy et al. 2000; Driscoll et al. 2001); changes in stream habitat due to habitat destruction, stream channelization, poor riparian management, or sedimentation (Curry and MacNeill 2004); and natural stochastic events (Roghair et al. 2002) have eliminated or severely reduced brook trout populations at local or regional scales (Bivens et al. 1985; SAMAB 1996; Galbreath et al. 2001; Habera et al. 2001;

McDougal et al. 2001). Although knowledge of these factors is important for watershed management, the limited budgets of most agencies would probably prevent the use of every factor in regional models or in determining how to subdivide larger regions into smaller ones for modeling purposes.

Identification of streams requiring management is difficult given past history and expectation of future changes. It may take over 50 years for the stream habitat to recover, even when past land use practices are remedied (Harding et al. 1998). Cases in point are subwatersheds that are predicted to contain intact or reduced populations but that are in fact sites of extirpation. The highest misclassifications were in cluster 6, an area where abusive land use practices historically caused extirpation of brook trout populations in subwatersheds (King 1937, 1939). Today, many of these subwatersheds are restored and protected (National Forest, National Park, and state lands) and have watershed attributes that would predict brook trout presence (i.e., high forest percentage, low percentage of agriculture use, and high elevation). However, as past land use practices abated and these subwatersheds recovered, stocking resulted in naturalization of rainbow trout (King 1937, 1939; Lennon 1967; Kelly et al. 1980). Naturalized rainbow trout now preclude the restoration of brook trout in these habitat-recovered subwatersheds. Unfortunately, available metrics for exotic fishes did not perform well in distinguishing between sites with differing brook trout status, probably because of a complex interaction among such factors as natural and manmade barriers, stocking history, and data set variability in identifying exotic fishes as threats at the subwatershed level. Exotic fishes like rainbow trout may have impacts at different scales throughout the brook trout's range (i.e., stream segment scale), and subwatershed-level analysis may be inappropriate for determining these effects. Some metrics may have greater influence on brook trout populations at different scales (Kocovsky and Carline 2006). For example, Rashleigh et al. (2005) were able to predict brook trout presence–absence in stream segments in the Mid-Atlantic Highlands with a correct classification rate of 79% using local-scale metrics of depth, temperature, substrate, percentage riffles, cover, and riparian vegetation. It is interesting that prediction rates were only slightly lower for our models, which were based on larger-scale data.

Despite obvious difficulties in predicting which sites to manage, we need evaluations of the integrity of watersheds over the native range of brook trout to guide decision makers, managers, and the public in setting priorities for watershed-level conservation, restoration, and monitoring programs. Improving our understanding of a species' current distribution and population status is one of the key tools in conservation (Williams et al. 1993; Warren et al. 1997; EBTJV 2006). The ability to predict site quality is essential for conservation and restoration. We believe that models based on a species' full data set, while useful, could miss important smaller-scale patterns.

Our models demonstrated the importance of elevation in all six clusters. The large values and negative signs of the coefficient estimates were indicative of lower stream temperature and a lesser degree of human disturbance at higher elevations. Even though elevation is not a metric that land managers can control, it suggests the importance of managing high-elevation streams for purposes of maintaining self-sustaining brook trout populations. Differential significance of the four remaining predictor variables and their estimated values in each cluster can aid land managers in setting priorities for protective management decisions.

Our models can be used in several ways. First, identification of a cluster that primarily contains sites of brook trout presence suggests a region where preservation and maintenance constitute the correct strategy. Second, areas with irregular or unexpected patterns of predicted brook trout status suggest that further investigation is required for better management decisions. Third, one can use well-performing predictive models for some study areas to predict future subwatershed status. Also, interpretation of the resulting logistic models can aid in managerial actions if combined with other professional knowledge (especially historical information).

Users of our method can easily modify the model performance criterion to achieve various research goals. We used the overall AFCCF as the criterion for obtaining better overall classification relative to the single-model approach. If one is interested in finding a "hot spot" within the region where the model can describe the stressor–response relationship well, the maximum AFCCF can be used as the criterion. Although we used latitude and longitude as clustering variables, other variables are possible. In another application on water quality in the Mid-Atlantic Highlands, we used elevation and stream width as the clustering variables. One can also easily extend this work to other application areas provided that (1) partitioning of the entire data set makes intuitive sense and (2) there are sensible clustering variables.

Alternative approaches are available for model-based clustering. Holmes et al. (1999) used Bayesian partitioning modeling to split a large region into disjoint subregions. Their examples involved regression models that assumed normal or multinomial responses. Denison and Holmes (2001) also extended

the method to count data analysis. Chipman et al. (2002) proposed Bayesian treed models (BTMs) as an extension of classification trees. The BTMs use a set of variables to split the data in a manner similar to that done in classification trees, but then a logistic or normal regression model is employed as the final step in each node of the tree. This method allows for richer models in each partition but permits only axis-parallel partitions, producing clusters of rectangular shape. Since each step of the partitioning process involves a binary cut, the resulting shapes are rather regular. Further, BTMs are not invariant to transformation of spatial coordinates. Hence, changing the shape of the spatial region could result in different clusters.

Acknowledgments

We thank the Eastern Brook Trout Joint Venture for the data set used in this analysis. A Science to Achieve Results Grant (Number RD-83136801–0) from the U.S. Environmental Protection Agency provided partial support. Feng Gao and David Farrar provided suggestions that improved the methodology and results. Finally, we are grateful to the associate editor for his patience and comments and to the reviewers for their helpful suggestions. Reference to trade names does not imply endorsement by the U.S. Government.

References

Banerjee, S. 2004. On geodetic distance computations in spatial modeling. Biometrics 61:617–625.

Belford, D. A., and W. R. Gould. 1989. An evaluation of trout passage through six highway culverts in Montana. North American Journal of Fisheries Management 9:437–445.

Bivens, R. D., R. J. Strange, and D. C. Peterson. 1985. Current distribution of the native brook trout in the Appalachian region of Tennessee. Journal of the Tennessee Academy of Science 60:101–105.

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

Clayton, J. L., E. S. Dannaway, R. Menendez, H. W. Rafah, J. J. Renton, S. M. Sherlock, and P. E. Zurbuch. 1998. Application of limestone to restore fish communities in acidified streams. North American Journal of Fisheries Management 18:347–360.

Curry, R. A., and W. S. MacNeill. 2004. Population-level responses to sediment during early life in brook trout. Journal of the North American Benthological Society 23:140–150.

Davis, W. S., and T. P. Simon. 1995. Biological assessment and criteria: tools for watershed resource planning and decision making. Lewis Publishers, Washington, D.C.

Denison, D. G. T., and C. C. Holmes. 2001. Bayesian partitioning for estimating disease risk. Biometrics 57:143–149.

Driscoll, C. T., G. B. Lawrence, A. J. Bulger, T. J. Butler, C. S. Cronan, C. Eagar, K. F. Lambert, G. E. Likens, J. L. Stoddard, and K. C. Weathers. 2001. Acidic deposition in the northeastern United States: sources and inputs, ecosystem effects, and management strategies. BioScience 51:180–198.

EBTJV (Eastern Brook Trout Joint Venture). 2006. Eastern Brook Trout Joint Venture Web site. International Association of Fish and Wildlife Agencies, Washington, D.C. Available: www.easternbrooktrout.org. (December 2006).

Fiss, F. C., and R. F. Carline. 1993. Survival of brook trout embryos in three episodically acidified streams. Transactions of the American Fisheries Society 122:268–278.

Frissell, C. A., and D. Bayles. 1996. Ecosystem management and the conservation of aquatic biodiversity and ecological integrity. Water Resources Bulletin 32:229–240.

Gagen, C. J., W. E. Sharpe, and R. F. Carline. 1993. Mortality of brook trout, mottled sculpins, and slimy sculpins during acidic episodes. Transactions of the American Fisheries Society 122:616–628.

Galbreath, P. F., N. D. Adams, S. Z. Guffey, C. J. Moore, and J. L. West. 2001. Persistence of native southern Appalachian brook trout populations in the Pigeon River system, North Carolina. North American Journal of Fisheries Management 21:927–934.

Gibson, R. J., R. L. Haedrich, and C. M. Wernerheim. 2005. Loss of fish habitat as a consequence of inappropriately constructed stream crossings. Fisheries 30(1):10–17.

Habera, J. W., R. J. Strange, and R. D. Bivens. 2001. A revised outlook for Tennessee's brook trout. Journal of the Tennessee Academy of Science 76:68–73.

Harding, J. S., E. F. Benfield, P. V. Bolstad, G. S. Helfman, and E. B. D. Jones III. 1998. Stream biodiversity: the ghost of land use past. Proceedings of the National Academy of Sciences of the USA 95:14843–14847.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer-Verlag, New York.

Holmes, C. C., D. G. T. Denison, and B. K. Mallick. 1999. Bayesian partitioning for classification and regression. Imperial College, Technical Report, London.

Hudy, M., D. M. Downey, and D. W. Bowman. 2000. Successful restoration of an acidified native brook trout stream through mitigation with limestone sand. North American Journal of Fisheries Management 20:453–466.

Hudy, M., T. M. Thieling, N. Gillespie, and E. P. Smith. 2006. Distribution, status, and perturbations to brook trout within the eastern United States. Final Report to the Eastern Brook Trout Joint Venture, International Association of Fish and Wildlife Agencies, Washington, D.C. Available: www.easternbrooktrout.org (December 2006).

Jenkins, R. E., and N. M. Burkhead. 1993. Freshwater fishes of Virginia. American Fisheries Society, Bethesda, Maryland.

Kelly, G. A., J. S. Griffith, and R. D. Jones. 1980. Changes in distribution of trout in Great Smoky Mountains National Park, 1900–1970. U.S. Fish and Wildlife Service Technical Papers 102.

King, W. 1937. Notes on the distribution of native speckled and rainbow trout in the streams of Great Smoky Mountains National Park. Journal of the Tennessee Academy of Science 12:351–361.

King, W. 1939. A program for the management of fish resources in Great Smoky Mountains National Park. Transactions of the American Fisheries Society 68:86–95.

Kocovsky, P. M., and R. F. Carline. 2006. Influence of landscape-scale factors in limiting brook trout populations in Pennsylvania streams. Transactions of the American Fisheries Society 135:76–88.

Larson, G. L., and S. E. Moore. 1985. Encroachment of exotic rainbow trout into stream populations of native brook trout in the southern Appalachian mountains. Transactions of the American Fisheries Society 114:195–203.

Lennon, R. E. 1967. Brook trout of Great Smoky Mountains National Park. U.S. Fish and Wildlife Service Technical Papers 15.

Li, K. C., and N. Duan. 1989. Regression analysis under link violation. Annals of Statistics 17:1009–1052.

MacCrimmon, H. R., and J. S. Campbell. 1969. World distribution of brook trout, *Salvelinus fontinalis*. Journal of the Fisheries Research Board of Canada 26:1699–1725.

Marschall, E. A., and L. B. Crowder. 1996. Assessing population responses to multiple anthropogenic effects: a case study with brook trout. Ecological Applications 6:152–167.

Master, L. L., S. R. Flack, and B. A. Stein, editors. 1998. Rivers of life: critical watersheds for protecting freshwater biodiversity. Nature Conservancy, Arlington, Virginia.

McDougal, L. A., K. M. Russell, and K. N. Leftwich, editors. 2001. A conservation assessment of freshwater fauna and habitat in the southern national forests. Report R8-TP 35, U.S. Forest Service, Southern Region, Report, Atlanta, Georgia.

Meisner, J. D. 1990. Effect of climatic warming on the southern margins of the native range of brook trout, *Salvelinus fontinalis*. Canadian Journal of Fisheries and Aquatic Sciences 47:1065–1070.

Møller, J. 1994. Lectures on random Voronoi tessellations. Lecture notes in statistics 87. Springer-Verlag, New York.

Moore, S. E., G. L. Larson, and B. Ridley. 1986. Population control of exotic rainbow trout in streams of a natural area park. Environmental Management 10:215–219.

Moore, S. E., B. Ridley, and G. L. Larson. 1983. Standing crops of brook trout concurrent with removal of rainbow trout from selected streams in Great Smoky Mountains National Park. North American Journal of Fisheries Management 3:72–80.

National Atmospheric Deposition Program. 2005. Isopleth grids. National Atmospheric Deposition Program. Available: nadp.sws.uiuc.edu. (July 2005).

Navtech. 2001. Navstreets: street data. CD-ROM, version 2.5. Navtech, Springfield, Virginia. Available: www.navtech.com.data/data.html. (March 2006).

Nichols, S., P. Sloane, J. Coysh, C. Williams, and R. Norris. 2000. Australian capital territory, Australian River Assessment System. Cooperative Research Center for Freshwater Ecology, University of Canberra, Australia.

Nislow, K. H., and W. H. Lowe. 2003. Influences of logging history and stream pH on brook trout abundance in first-order streams in New Hampshire. Transactions of the American Fisheries Society 132:166–171.

Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu. 2000. Spatial tessellations: concepts and applications of Voronoi diagrams, 2nd edition. Wiley, New York.

Pepe, M. S., T. Cai, and G. Longton. 2005 Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics 62:221–229.

Rashleigh, B., R. P. Parmar, J. M. Johnston, and M. C. Barber. 2005. Predictive habitat models for the occurrence of stream fishes in the Mid-Atlantic Highlands. North American Journal of Fisheries Management 25:1353–1336.

Rieman, B. E., D. C. Lee, and R. F. Thurow. 1997. Distribution, status, and likely future trends of bull trout within the Columbia River and Klamath River basins. North American Journal of Fisheries Management 17:1111–1125.

Roghair, C. N., C. A. Dolloff, and M. K. Underwood. 2002. Response of a brook trout population and instream habitat to a catastrophic flood and debris flow. Transactions of the American Fisheries Society 131:718–730.

SAMAB (Southern Appalachian Man and the Biosphere). 1996. The southern Appalachian assessment atmospheric technical report. Report 3 of 5. U.S. Forest Service, Southern Region, Atlanta.

Strange, R. J., and J. W. Habera. 1998. No net loss of brook trout distribution in areas of sympatry with rainbow trout in Tennessee streams. Transactions of the American Fisheries Society 127:434–440.

Thieling, T. M. 2006. Assessment and predictive model for brook trout (*Salvelinus fontinalis*) population status in the eastern United States. Master's thesis. James Madison University, Harrisonburg, Virginia.

Thurow, R. F., D. C. Lee, and B. E. Reiman. 1997. Distribution and status of seven native salmonids in the interior Columbia River basin and portions of the Klamath River and Great basins. North American Journal of Fisheries Management 17:1094–1110.

USACE (U.S. Army Corps of Engineers). 1998. National inventory of dams data. USACE, Washington, D.C. Available: crunch.tec.army.mil. (November 2004).

U.S. Census Bureau. 2002. Census 2000: summary files. U.S. Census Bureau. Available: www.census.gov. (November 2002).

USGS (U.S. Geological Survey). 2004. National land cover dataset (NLCD 1992). USGS, Reston, Virginia. Available: landcover.usgs.govnatllandcover.php. (June 2004).

Warren, M. L., Jr., P. L. Angermeier, B. M. Burr, and W. R. Haag. 1997. Decline of a diverse fish fauna: patterns of imperilment and protection in the southeastern United States. Pages 105–164 *in* G. W. Benz and D. E. Collins, editors. Aquatic fauna in peril: the southeastern perspective. Southeast Aquatic Research Institute, Special Publication 1, Lenz Design and Communications, Decatur, Georgia.

Wilkinson, L. 1999. SYSTAT, version 9. SPSS, Chicago.

Williams, J. D., M. L. Warren, Jr., K. S. Cummings, J. L. Harris, and R. J. Neves. 1993. Conservation status of freshwater mussels of the United States and Canada. Fisheries 18(9):6–22.

Wright, J. F., D. Moss, P. D. Armitage, and M. T. Furse. 1984. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. Freshwater Biology 14:221–256.

Yarnell, S. L. 1998. The southern Appalachians: a history of the landscape. U.S. Forest Service General Technical Report SRS-18.