An evaluation of parametric and nonparametric models of fish population response

Submitted by:

James T. Peterson

U.S. Department of Agriculture Forest Service,
Rocky Mountain Research Station, Boise, Idaho

Prepared for:

U.S. Department of Energy
Bonneville Power Administration
Environment, Fish and Wildlife
P.O. Box 3621
Portland, OR 97208-3621

Project Number 92-032-00
Contract Number 92AI25866

November 1999

An Evaluation of Parametric and Nonparametric Models of Fish Population Response

Timothy C. Haas[1], James T. Peterson[2,3], and Danny C. Lee[2,4]

[1]School of Business Administration, University of Wisconsin at Milwaukee, P.0. Box 742, Milwaukee, WI 53201, USA

[2]U.S.D.A. Forest Service, Rocky Mountain Research Station, 316 East Myrtle Street, Boise, Idaho 83702 USA

[3] Corresponding author. Current address: Georgia Cooperative Fish and Wildlife Research Unit, Warnell School of Forest Resources, University of Georgia, Athens, GA 30602 USA, peterson@smokey.forestry.uga.edu.

[4] Current address: USDA Forest Service, Sierra Nevada Conservation Framework, 801 I Street, Sacramento, CA 95814, USA

Abstract

Predicting the distribution or status of animal populations at large scales often requires the use of broad-scale information describing landforms, climate, vegetation, etc. These data, however, often consist of mixtures of continuous and categorical covariates and nonmultiplicative interactions among covariates, complicating statistical analyses. Using data from the interior Columbia River Basin, USA, we compared four methods for predicting the distribution of seven salmonid taxa using landscape information. Subwatersheds (mean size, 7800 ha) were characterized using a set of 12 covariates describing physiography, vegetation, and current land-use. The techniques included generalized logit modeling, classification trees, a nearest neighbor technique, and a modular neural network. We evaluated model performance using out-of-sample prediction accuracy via leave-one-out cross-validation and introduce a computer-intensive Monte Carlo hypothesis testing approach for examining the statistical significance of landscape covariates with the non-parametric methods. We found the modular neural network and the nearest-neighbor techniques to be the most accurate, but were difficult to summarize in ways that provided ecological insight. The modular neural network also required the most extensive computer resources for model fitting and hypothesis testing. The generalized logit models were readily interpretable, but were the least accurate– possibly due to nonlinear relationships and nonmultiplicative interactions among covariates. Substantial overlap among the statistically significant ($P<0.05$) covariates for each method suggested that each is capable of detecting similar relationships between responses and covariates. Consequently, we believe that employing one or more methods may provide greater biological insight without sacrificing prediction accuracy.

1. Introduction

Ecosystem management requires understanding of both fine-grained and coarse-grained ecological features (Levin, 1992). While remote data-capture technologies allow compilation of coarse-grained information describing landforms, climate, vegetation, etc., these technologies cannot detect fine-grained features such as the distribution of vertebrate species. Consequently, ecologists are increasingly relying upon models to predict the distribution or status of vertebrate populations (Miller et al., 1989; Kruse et al., 1997; Carroll et al., 1999; Naugle et al., 1999) and to examine the effect of environmental or anthropogenic impacts on those populations over large scales (Lee et al., 1997; Baxter et al., 1999; Dunham and Rieman, 1999). The accuracy of these predictions depends, in part, on the development of rigorous statistical models that relate environmental data, which often consisting of a mix of continuous (hereafter, quantitative covariates) and discrete-valued variables, to categorical population responses (e.g., species presence/absence). In these instances, the use of traditional parametric modeling techniques (e.g., linear regression) is inappropriate. Recent advances in the statistical and computing sciences have led to the development of sophisticated nonparametric methods for the analysis of these complex data sets (e.g., Lek and Guegan, 1999). However, statistical models and hypothesis tests for these techniques are not as well developed as traditional parametric approaches.

For purposes of both finding the best performing nonparametric model and for assessing the importance (significance) of a subset of covariates, some method of judging whether two nonparametric classifiers are significantly different in predictive performance is needed. Here we present such an approach and compare the predictive performance (accuracy) of 4 methods: generalized logit modeling, classification trees, a nearest neighbor technique, and modular neural network using existing data on landscape features and salmonid populations in the Northwestern U.S. Our goal was to illustrate and discuss alternative methods of modeling categorical responses, compare their predictive performance, and provide a means to identify important relationships between population responses and landscape covariates. These methods were selected for evaluation because they each represent recent statistical research on the classification problem. As an aside, linear discriminant analysis was not evaluated because it cannot incorporate qualitative

covariates (Johnson and Wichern, 1992) and is generally inferior to generalized logit modeling (Press and Wilson 1978).

1.2 Data description

We evaluated the performance of the 4 parametric and nonparametric methods using landscape and population status data for 7 salmonid taxa in the interior Columbia River Basin, USA. These data are thoroughly described in previous works (Lee et al., 1997; Rieman et al., 1997; Thurow et al., 1997) and are briefly reviewed here. The study area included the entire Columbia River Basin east of the Cascade Mountains and small portions of the Klamath and Great Basins, USA, which encompasses 58.6 million hectares in Idaho, Montana, Nevada, Oregon, Washington, and Wyoming, USA. Within the study area, subwatersheds (U.S. Geological Survey 6th code hydrologic units; mean size 7800 ha) were used as the basic unit for our analysis. Landscape data for each subwatershed were obtained from the Interior Columbia River Basin Project (Quigley and Arbelbide, 1997) and included 11 quantitative attributes and 1 qualitative (categorical) attribute with 10 levels (Table 1). These data represented various landform, climate, vegetation, and land use characteristics known to influence the structure and stability of lotic habitats and, presumably, fish populations.

The salmonid taxa analyzed were both ecologically and economically important to the region (Lee et al., 1997) and included bull trout (*Salvelinus confluentus*), redband trout (*Oncorhynchus mykiss gibbsi*), westslope cutthroat trout (*Oncorhynchus clarki lewisi*), Yellowstone cutthroat trout (*Oncorhynchus clarki bouvieri*), chinook salmon (*Oncorhynchus tshawytscha*), and steelhead (*Oncorhynchus mykiss mykiss*). Because life history requirements are likely to differ among anadromous stocks with very different migratory patterns (Thurow et al., 1997), chinook salmon were subdivided into 2 groups (stocks) based on Healey's (1991) definitions: *ocean-type*, those that migrate seaward as subyearlings and *stream-type*, those that migrate after rearing one or more years in freshwater. Using current (post 1993) empirical data, more than 150 governmental, tribal, and privately employed biologists classified the population status of each salmonid taxon in individual subwatersheds as *strong*– all major life history types are present and numbers are stable or increasing, *depressed*– one or more life history types absent or numbers decreasing, *absent*, or *unknown*. For the anadromous salmonids– chinook

salmon and steelhead, *migrant* status was used to classify subwatersheds functioning primarily as migration corridors (Rieman et al., 1997). Subwatersheds for which the population status was *unknown* were not included in the analyses. Note that some of these responses (*strong* and *depressed*) are the result of a judgment by a biologist and not a direct measurement; hence there potentially was a variability to the data that could not be explainable by the covariates.

2. Statistical modelling

2.1 Assessing predictive ability

The expected error rate (EER) of a classifier is the average error rate of the classifier averaged over all possible patterns of responses at the design points (i.e., covariate locations) (Lachenbruch 1975). To compute the EER, the distribution of the categorical response would need to be known at each design point. Then, an expected error rate could be computed over all possible such patterns at these design points. Note that EER is not the error rate of the classifier built from the particular sample in-hand (called the actual error rate) but rather the average over all possible response patterns at the design points - not just the one observed. Thus, the EER could provide the basis for comparing the performance of the various models of fish population response. EER also is a deterministic function of the parameters defining the response variable's distribution and hence can be thought of as a parameter itself.

Fukunaga and Kessel (1971) found the cross-validation estimator to be nearly unbiased and in fact, slightly conservative when used to estimate EER of a nonparametric estimator. Similarly, Efron (1983) concluded that with a large sample size, cross-validation gives a "nearly-unbiased" measure of overall predictive ability without excessive variance. Efron (1983) also concludes that although bootstrap estimates of predictive performance usually have low variance, they can also exhibit large negative bias - particularly when used with overfitted models. This last cautionary note concerning the use of a bootstrap estimate is echoed in a review paper on discrimination analysis: "However, in utterly nonparametric situations, the bootstrap can badly underestimate the misclassification costs (Breiman et al., 1984) and even be inconsistent." (Gnanadesikan et al.,1989). An alternative estimator of EER that uses the double bootstrap (Efron, 1983) may offer a lower variance estimate of EER that also has

small bias. However, numerical experimentation and comparison with cross-validation error rate estimates are needed to assess this alternative estimator's performance.

We estimated the EER (hereafter, $\hat{EER}$) via leave-one-out cross-validation as originally proposed by Lachenbruch (1965). Each observation in a data set was temporarily held out of the sample and the model fitted with the remaining n - 1 observations. The held-out observation was then predicted by the mode of this fitted model evaluated at the held-out observation's vector of covariate values. The average number of misclassifications were then summed over all observations and also over observations on each response category. For logit modeling, we also estimated the within-sample error rate, EER_w (also known as the apparent error rate), which was calculated by applying the model to the observations that were used during model fitting. Note that EER_w, is known to be a negatively biased (optimistic) estimator for the EER (Johnson and Wichern, 1992), but provides a relatively quick estimate of model performance when examining several complex models with large data sets, such as those used here.

2.2 Generalized logit model

2.2.1. Theory and definitions

Say that n multivariate observations on J multinomial random variables are taken. Let g be the number of populations or groups (unique combinations of covariates). For the i^{th} observation, let $\mathbf{Y}(x_i) = (Y_1(x_i),...,Y_J(x_i))'$ where $Y_j(\mathbf{x}_i)$ is the number of response category j occurrences in n_i trials. Define π_{ij} to be the probability that the j^{th} response category occurs on any particular trial. Then, $\mathbf{Y}(\mathbf{x}_i) \sim$ Multinomial(n_i, $\boldsymbol{\pi}_i$). Hence, at covariate vector $\mathbf{x}_i$, there is a pattern of response probabilities. Let the observed value of $Y_j(\mathbf{x}_i)$, y_{ij} be the number of response j occurrences observed on the i^{th} population.

A natural extension of linear model theory leads to a model of transformed response probabilities being linear in the covariate parameters. This is the *generalized logit model* (Agresti 1990). The j^{th} logit is:

$$\boldsymbol{\eta}_j(\mathbf{x}_i) = g\left(E\left[Y_j(\mathbf{x}_i)\right]\right) = g\left(\mu_j(\mathbf{x}_i)\right) \equiv \log\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \mathbf{x}_i'\boldsymbol{\beta}_j, \tag{1}$$

for $j = 1 \ldots, J-1$ baseline-category logits. The function $g(.)$ is called the link function between the observed variable's expected value in the original-scale space and its value in the transformed space. The multivariate linear model of the response probabilities is:

$$\begin{bmatrix} \eta_1(\mathbf{x}_1) \ldots \eta_{J-1}(\mathbf{x}_1) \\ \vdots \quad \vdots \quad \vdots \\ \eta_1(\mathbf{x}_n) \ldots \eta_{J-1}(\mathbf{x}_g) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{bmatrix} \beta_{1,0} \ldots \beta_{J-1,0} \\ \vdots \quad \vdots \quad \vdots \\ \beta_{1,p} \ldots \beta_{J-1,p} \end{bmatrix}. \tag{2}$$

For computational purposes, this model is written as:

$$\begin{bmatrix} \eta_1(\mathbf{x}_1) \\ \vdots \\ \eta_{J-1}(\mathbf{x}_1) \\ \vdots \\ \eta_1(\mathbf{x}_n) \\ \vdots \\ \eta_{J-1}(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} \begin{bmatrix} \beta_{1,0} \\ \vdots \\ \beta_{1,p} \\ \vdots \\ \beta_{J-1,0} \\ \vdots \\ \beta_{J-1,p} \end{bmatrix} \tag{3}$$

where

$$Z_i \equiv \begin{bmatrix} \mathbf{x}_i' & \mathbf{0}' & \ldots & \ldots & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_i' & \mathbf{0}' & \ldots & \mathbf{0}' & \mathbf{0}' \\ & & \vdots & & & \\ \mathbf{0}' & \ldots & & & \mathbf{0}' & \mathbf{x}_i' \end{bmatrix}. \tag{4}$$

In vector notation,

$$\begin{bmatrix} \boldsymbol{\eta}(\mathbf{x}_1) \\ \vdots \\ \boldsymbol{\eta}(\mathbf{x}_n) \end{bmatrix} = Z\boldsymbol{\beta}. \tag{5}$$

The j^{th} response category probability is a nonlinear function of the parameter vector, $\boldsymbol{\beta}_j$:

$$\pi_{ij} = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_j\right)}{\sum_{k=1}^{J}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_k\right)}, \tag{6}$$

For identifiably, $\boldsymbol{\beta}_j \equiv 0$ (and hence $\eta_J(.) = 0$) and the response category probabilities can be recovered from the logits via the *response function*, h(.):

$$\pi_{ij} = h_j\left(\boldsymbol{\eta}(\mathbf{x}_i)\right) = \frac{\exp\left(\eta(\mathbf{x}_i)\right)}{1+\sum_{k=1}^{J-1}\exp\left(\eta_k(\mathbf{x}_i)\right)}, \tag{7}$$

for $j = 1,...., J-1$.

The joint probability of observing the data is:

$$\prod_{i=1}^{g} P\left(Y_1(\mathbf{x}_i) = y_{i1},\ldots,Y_J(\mathbf{x}_i) = y_{iJ}\right) = (\text{a constant}) \prod_{i=1}^{g}\left[\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\right]\left[1-\sum_{j=1}^{J-1}\pi_{ij}\right]^{n_i-\sum_{j=1}^{J-1}y_{ij}}, \tag{8}$$

and hence the log-likelihood function is $l = \sum_{i=1}^{g}\sum_{j=1}^{J-1} y_{ij}\eta_{ij} - n_i\log\left\{\sum_{k=1}^{J}\exp(\eta_k(\mathbf{x}_i))\right\}$.

The vector $\boldsymbol{\beta}$ that maximizes this function is the estimate of the model's parameters.

A well-known form of the generalized logit model, binary logistic regression, is used when there are 2 response categories (i.e., $J = 2$). Here, there is only one logit called the log-odds of observing the first category from the i^{th} population: $\log(\pi_{i1}/\pi_{i2}) = \log(\pi_{i1}/(1-\pi_{i1}))$.

2.1.2 Cumulative logit model

If the response categories are ordered, a reduction in the number of model parameters relative to the baseline-category logit model can be achieved by incorporating this assumption into the model. One way to do this is with cumulative logits:

$$\eta_{ij} = \log\left(\sum_{k=1}^{j}\pi_{ik} /(1-\sum_{k=1}^{j}\pi_{ik})\right) \tag{9}$$

$$= \alpha_j + \mathbf{x}_i'\boldsymbol{\beta},\ j = 1,...,J-1 \tag{10}$$

(Agresti 1990). Note that for $J > 2$, the logit models share a common set of covariate effect parameters ($\boldsymbol{\beta}$) and differ only in their intercept parameters (α_j's). The assumption is that differences among the logits are due only to an order-driven shift in the overall

mean. Hence, these models form $J-1$ parallel lines (for $J > 2$). A test of this parallel lines assumption is reviewed below.

2.1.3 Hypothesis testing and model selection

Let $\Sigma_i(\boldsymbol{\beta})$ be the covariance matrix of $\mathbf{Y}(\mathbf{x}_i)$ and $D_i(\boldsymbol{\beta})$ be the Jacobian of the response function evaluated at $\boldsymbol{\beta}$.

$$D_i \equiv \begin{bmatrix} \partial h(\eta_1)/\partial\eta_1 & \cdots & \partial h(\eta_1)/\partial\eta_q \\ \vdots & \vdots & \vdots \\ \partial h(\eta_q)/\partial\eta_1 & \cdots & \partial h(\eta_q)/\partial\eta_q \end{bmatrix}. \tag{11}$$

Let $W_i(\boldsymbol{\beta}) = D_i(\boldsymbol{\beta})\Sigma_i(\boldsymbol{\beta})^{-1}(\boldsymbol{\beta})\mathrm{D}_i(\boldsymbol{\beta})'$ be an approximation to$\mathrm{Cov}^{-1}[g(\mathbf{Y}(\mathbf{x}_i)]$. The gradient of the log-likelihood is called the *score function*:

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) &\equiv \partial l/\partial\boldsymbol{\beta} \\ &= \sum_{i=1}^{g} Z' D_i(\boldsymbol{\beta}) \left[\frac{\Sigma_i}{n_i}\right]^{-1} (\boldsymbol{\beta})\left[\overline{\mathbf{y}}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right] \end{aligned} \tag{12}$$

and the negative expected value of the log-likelihood's Hessian matrix is called the *Fisher Information matrix*:

$$\begin{aligned} F(\boldsymbol{\beta}) &\equiv -E\left[\partial^2 l/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right] \\ &= \sum_{i=1}^{g} Z_i' W_i(\boldsymbol{\beta}) Z_i \end{aligned} \quad . \tag{13}$$

Consider the case of adding one covariate to the model. If the parameters corresponding to the added covariate are contained in the vector $\boldsymbol{\beta}_2$ and those of the original model in $\boldsymbol{\beta}_1$, then $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$. Partition s and F with respect to $\boldsymbol{\beta}$:

$$s = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, \quad F = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}. \tag{14}$$

Let the null hypothesis be H_o: $\boldsymbol{\beta}_2 = 0$ and let $\tilde{s}$ and $\tilde{F}$ be s and F evaluated at the parameter estimates under H_o. Then, letting $A \equiv F - 1$, the *score statistic* is $u \equiv \tilde{s}_2' \tilde{A}_{22} \tilde{s}_2$ and is asymptotically distributed as $\chi^2_{dim(\hat{\boldsymbol{\alpha}}_2)}$ (Fahrmeir and Tutz, 1994). An important property of the score statistic is that convergence to its asymptotic distribution is not a function of the individual cell counts, i.e., convergence is for $n \to \infty$ whether or not n_i is also increasing (Fahrmeir, 1988; Fahrmeir and Tutz, 1994; Fahrmeir and Kaufmann, 1987). Essentially, the asymptotic distribution depends on the design points being

sufficiently dispersed throughout the design space (covariate space) but replication at design points, unlike the power-divergence statistic (below), is not required.

A related statistic, useful for testing a reduced model after a full model has been successfully fitted (i.e., backward elimination), is the Wald statistic: $w \equiv \hat{\boldsymbol{\beta}}_2' \hat{A}_{22}^{-1} \hat{\boldsymbol{\beta}}_2$ where $\hat{\boldsymbol{\beta}}_2$ and $\hat{F}$ are estimated under the full model (Fahrmeir and Tutz, 1994). The above remarks on the asymptotic distribution of the score statistic also apply to the Wald statistic.

Fahrmeir and Tutz (1994) recommend forward selection as a workable model selection procedure when the fitting of a complete model would fail due to a large number of covariates. For example, if there are 20 quantitative covariates, the full two-way interaction model would contain 1 + 20 + 20(20 - 1)/2 = 211 terms making it infeasible to start the model with all two-way interactions included and perform (say) backward selection. Forward selection proceeds by computing *u* for each covariate (or interaction) not already in the model and selecting for inclusion that variable (or interaction) that gives the largest value of *u* that is also larger than a critical value defined by a modeler-selected Type I error probability, α. There is no guarantee however, that the forward selection method will find an adequate model for the data since the method not only starts with an inadequate model but then proceeds to consider only variables or interactions one at a time (Christensen, 1990). One way to address this drawback (and one approach we took) is to start with a model that contains all of the main effects and then forward-select two-way interactions. This approach is best suited when a large data set is available so that the loss of power due to having (perhaps) too many covariates in the smallest model is not too great.

2.1.4 Goodness-of-fit and residual analysis

Power divergence statistics

For λ a real number, the class of *power-divergence* goodness-of-fit statistics is defined to be $D_\lambda = \sum_{i=1}^{g} \frac{2n_i}{\lambda(\lambda+1)} \sum_{j=1}^{J} \bar{y}_{ij} \left[(\bar{y}_{ij} / \hat{\pi}_{ij})^{\lambda} - 1 \right]$ where $\bar{y}_{ij} \equiv y_{ij}/n_i$ (Fahrmeir and Tutz, 1994). Letting $\lambda \to \infty$ gives the *deviance*, $D_0 \equiv \sum_{i=1}^{g} n_i \sum_{j=1}^{J} \bar{y}_{ij} \log(\bar{y}_{ij} / \hat{\pi}_{ij})$ with the understanding that $\bar{y}_{ij} \log(\bar{y}_{ij} / \hat{\pi}_{ij})$ is 0 if $\bar{y}_{ij} = 0$ (Fahrmeir and Tutz, 1994).

Asymptotic analysis shows that $D_0 \overset{a}{\sim} \chi^2_{(g(J-1)-(p+1))}$ if $n_i \to \infty$ for all i, and g is fixed. Hence, when this *increasing-cell-counts* asymptotic distribution is approximated, D_0 can be used as a goodness-of-fit statistic. The value $\lambda = 1$ gives the *Pearson Goodness-of-Fit* statistic, $D_1 \equiv \sum_{i=1}^{g} n_i \sum_{j=1}^{J} (\bar{y}_{ij} - \hat{\pi}_{ij})^2 / \hat{\pi}_{ij}$ (Fahrmeir and Tutz, 1994) and has the same increasing-cell-counts asymptotic distribution as D_0.

Pearson residuals are $\mathbf{r}_i \equiv \sum_i^{-1/2} (\bar{\mathbf{y}}_i - \hat{\pi}_i)$, $i = 1,...g$ and the *studentized* Pearson residuals are $\mathbf{r}_i^{(s)} \equiv (I - H_{ii})^{-1/2} \mathbf{r}_i$ where $H_{ii} \equiv W_i^{T/2} Z_i F^{-1} Z_i^T W_i^{1/2}$ is the generalized hat matrix (Fahrmeir and Tutz, 1994). When the asymptotic χ^2 distribution for D_0 and D_1 is well-approximated, $\mathbf{r}^{(s)'}\mathbf{r}^{(s)}$, $i = 1,..., g$ will be approximately i.i.d. $\chi^2_{(g(J-1)-(p+1))/g}$ (Fahrmeir and Tutz, 1994) and hence a $\chi^2_{(g(J-1)-(p+1))/g}$ probability plot of $\mathbf{r}^{(s)'}\mathbf{r}^{(s)}$, $i = 1,..., g$ should be approximately one-to-one.

If cell counts are small and/or diagnostic plots suggest that the increasing-cell-counts asymptotic distribution of the test statistic is not holding, then *increasing-cells* asymptotics may be more appropriate. These are discussed next.

Increasing-cells asymptotics

If n_i is small and bounded but the number of populations, g is increasing, then, under certain conditions, $(D_1 - \mu_1)/\sigma_1 \overset{a}{\sim} N(0,1)$ (Osius and Rojek, 1992). These conditions are: (1) p_{ij} is bounded away from zero as $g \to \infty$, (2) the quantity g/σ_1^2 (Osius and Rojek, 1992) note that if the covariates are bounded, condition 1 is always satisfied. These authors show that the asymptotic mean, μ_1 is $g(J - 1)$ and the asymptotic variance is:

$$\sigma_1^2 = 2g(J-1) + \sum_{i=1}^{g} \frac{1}{n_i} \left[\left(\sum_{j=1}^{J} 1/\pi_{ij} \right) - J^2 - 2(J-1) \right] - \mathbf{c}_1' F^{-1} \mathbf{c}_1, \qquad (15)$$

where

$$\mathbf{c}_1 = \sum_{i=1}^{g}\sum_{j=1}^{J}(1/\pi_{ij})\begin{bmatrix}\partial\pi_{ij}/\partial\beta_0 \\ \vdots \\ \partial\pi_{ij}/\partial\beta_p\end{bmatrix}, \qquad (16)$$

evaluated at $\hat{\pi}_{ij}$. When the cell sizes are all exactly 1, as was the case with the salmonid population response data, Osius and Rojek (1992) recommend a two-sided test.

Consequences of failed asymptotics

If cell counts are small, the median of D_0 and D_1's true distribution will be less than the increasing-cell-count asymptotic χ^2 distribution's median. The result of this is that using the χ^2 distribution to find the critical value for rejecting the null hypothesis, H_0: "correct model," will give a much larger value than the true distribution's value for a fixed α. Hence, the goodness-of-fit test will rarely reject, i.e., poorly fitting models will rarely be detected (Agresti, 1990; Haberman 1988). There is no clear rule for how large cell counts need to be for reasonable approximation of the increasing-cell-counts asymptotic distribution. One well-known heuristic is Cochran's (1954) recommendation that no more than 20% of the cells have < 5 observations. Agresti (1990) states that in this case, goodness-of-fit statistics are not appropriate for testing the fit of a model but can be used to compare models. The same author also states however, that chi-squared goodness-of-fit statistics can be completely uninformative for highly sparse data.

Hence, when using a power-divergence statistic with small cell counts for purposes of assessing model goodness-of-fit, one strategy is to first determine if the χ^2 asymptotic distribution is well-approximated and if so, use increasing-cell-counts asymptotics to test hypotheses. Apparently, diagnostics (e.g., plots) do not exist for assessing how well the increasing-cells asymptotic distribution of the D_1 statistic is being approximated by the observed process. Note that it is misleading to examine a normal quantile-quantile (Q-Q) plot of the Osius and Rojek-standardized residuals because the derivations of the asymptotic normality of the test statistic, unlike those under increasing-cell-count asymptotics, do not begin with the individual residuals (see Agresti, 1990 and Osius and Rojek, 1992). Because of this difficulty in assessing the appropriateness of increasing-cells asymptotics, an *omnibus* test should always be computed whenever

increasing-cell-count asymptotics is in doubt. Andrews (1988) provides theoretical justification for a general family of such tests that have a χ^2 distribution when the model-based distribution and the data-generating distribution are the same. Such a test is computed by first grouping the observations into a fixed number of nearly homogeneous groups and then conducting a modified χ^2 test of the similarity between two vectors of counts. Because of the grouping step, under the null hypothesis, the test statistic's asymptotic distribution is always approached as the total sample size increases (described below). It should be noted however, that for practical purposes, the most relevant measure of model goodness-of-fit is an estimate of EER.

Test of parallel lines assumption

For the cumulative logit model, if parallel lines is not assumed (full model), then $\eta_{ij} = \alpha_j + \mathbf{x}_i'\beta_j$, $j = 1, \ldots, J$ - I where $\boldsymbol{\beta}_j$ equals $\boldsymbol{\beta}_1$ for $j = 1$ and equals $\boldsymbol{\beta}_1 + \gamma_j$ for $j > 1$. Here, γ_j represents the differences between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_j$, $j > 1$. Under the reduced model, H_0: $\gamma_2 = \ldots \gamma_{J-1} = 0$. By partitioning s and F so that the second partition contains $\gamma_2, \ldots \gamma_{J-1}$, a score statistic can be computed that is asymptotically chi-squared under H_0 (SAS Institute, 1989).

Hosmer-Lemeshow test

For the case of $J = 2$ (binary observations), let π_i be the probability of $Y_i = 1$ (referred to as the event) for the i^{th} population. Sort the observations by $\hat{\pi}_i$ and then partition this sorted list into $h = 10$ groups of nearly equal size. Let the k^{th} group's size be n_k. Let $\overline{\pi}_i$ be the average estimated event probability and o_k the observed frequency of events within group k. Hosmer and Lemeshow (1980) show through a simulation study that if the model used to compute $\hat{\pi}$ is true, the statistic:

$$HL = \sum_{k=1}^{h} \frac{(o_k - n_k \overline{\pi}_k)^2}{(n_k \overline{\pi}_k (1 - \overline{\pi}_k))}, \tag{17}$$

has approximately a χ^2 distribution as the overall sample size, n becomes large. Hosmer and Lemeshow (1980) do not discuss how the observations should be sorted for the case of $J > 2$.

Andrews chi-square test

The Andrews χ^2 test is a generalization of the Hosmer-Lemeshow test and is valid for any number of responses (J). The procedure for computing this test for a generalized logit model with all $n_i = 1$ is as follows.

Step 1: Use $\hat{p}_{i1},\ldots,\hat{p}_{iq}$ as variables for partitioning the data into K clusters using K-Means clustering (Johnson and Wichern, 1992).

Step 2: Regard $h = KJ$ as the number of groups or cells for the test. Let $y_{ij} = 1$ if the j^{th} response occurs on the i^{th} observation and 0 otherwise. Form the vector $\hat{\boldsymbol{\Gamma}}_i = (\boldsymbol{I}'_{i1},\ldots,\boldsymbol{I}'_{iK})'$ where

$\boldsymbol{I}_{i1} = \left(I_{\{1,C_l\}}(y_{i1},\hat{\mathbf{p}}_i), I_{\{2,C_l\}}(y_{i2},\hat{\mathbf{p}}_i),\ldots,I_{\{J,C_l\}}(y_{iJ},\hat{\mathbf{p}}_i)\right)'$ and C_l, is the l^{th} cluster. The $\hat{\boldsymbol{\Gamma}}_i$ is a vector of indicator variables. For the i^{th} observation, the indicator variable corresponding to the j^{th} response category and k^{th} cluster will equal one if the j^{th} response is observed and $\hat{\mathbf{p}}_i$ belongs to the k^{th} cluster.

Also form the vector $\hat{\boldsymbol{F}}_i = \left(\hat{\mathbf{p}}'_i I_{\{C_1\}}(\hat{\mathbf{p}}_i),\ldots,\hat{\mathbf{p}}'_i I_{\{C_K\}}(\hat{\mathbf{p}}_i)\right)'$. Hence, $\boldsymbol{F}_i$ is similar to $\hat{\boldsymbol{\Gamma}}_i$ except the binary response vector for the observed partition is replaced with the vector of model-based probabilities.

Finally, form the matrix H with the i^{th} row defined to be $\left(\hat{\boldsymbol{\Gamma}}_i - \hat{\boldsymbol{F}}'_i, \mathbf{s}'_i\right)$ where $\mathbf{s}'_i = \partial\, log\ f(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\beta}) / \partial\boldsymbol{\beta}$, the score function for the i^{th} observation.

Step 3: Letting **one** be the n x 1 vector of all 1's, compute the test statistic $X^2(\hat{\Gamma},\hat{\beta}) = \mathbf{1}'H(H'H)^{-}H'\mathbf{1}$ and compare its value to the chi-square distribution with $h - K$ degrees of freedom.

2.2 Classification Trees

2.2.1 Theory and definitions

Tree-based classification is one of a larger set of techniques recently developed for analyzing non-standard data (e.g., mixtures of quantitative and qualitative covariates; Brieman et al., 1984). Classification trees (hereafter simply tree) consist of a collection of binary decision rules called nodes, connected by directed arcs (Fig. 1), and created during

a procedure known as recursive partitioning (described below). At a node, the value of a particular covariate (x_o) determines which of the two arcs is followed. If x_o is a quantitative covariate, the rule is: follow left arc if $x_o \leq \alpha$, otherwise, follow right arc. If x_o is qualitative, the rule is: follow left arc if $x_o \in A$ otherwise, follow right arc. Terminal nodes are the model's predicted values of the response category. The structure of tree classification rules differ significantly from techniques, such as discriminant analysis and generalized logit models, where classification rules are based on linear combinations of covariates. This makes tree classifiers more flexible than traditional linear methods. For instance, tree models can incorporate qualitative covariates with more than 2 levels, integrate complex mixtures of data types, and automatically incorporate interactions among covariates. Tree models also do not necessarily use all covariates and may use some covariates more than once.

Mathematically, a tree can be represented as a set of nodes. After the tree has been fitted, each node has been assigned a subset of the observations. These subsets then can be used to estimate the response variable's probability distribution at each terminal node. This distribution is the empirical distribution of observations "falling" to that terminal node. When a new covariate vector is "dropped" down the tree, this observed distribution serves as the tree's prediction of the response variable (Fig. 1). Typically, the mode of this distribution is the predicted response category and terminal node's label.

2.2.2 Tree model fitting

Binary trees are created by repeatedly splitting the data set into 2 smaller subsets using binary rule-sets. The tree growing process begins with all of the data at a single location known as a parent or root node (e.g., t_1 in Fig. 1). This node is split into two child nodes (e.g., t_2 and t_3 in Fig. 1) using a rule generated during recursive partitioning. The recursive partitioning process searches for a covariate and its cutoff value (partitioning rule) that results in the greatest within-partition homogeneity for the response categories' distribution. In other words, the data is split into two subsets, each containing greater proportions of one response category. This covariate and partitioning rule defines the root node. This process is continued recursively down each “branch” of the tree until the size of a partition at any node is smaller than a prespecified stopping value (i.e., the minimum partition size).

Several measures of within-partition homogeneity are possible (Breiman et al., 1984). For the current study we used the deviance, defined to be the negative of twice the log-likelihood (Chambers and Hastie, 1992). The partitioning rule at node *t* is found by exhaustively searching for a covariate and partitioning pair that yields the largest reduction in node *t*'s deviance. The reduction in deviance at a particular root node *t* is estimated as:

$$2\sum_{k=1}^{allclasses}\left[n_{lk}\,log\left\{\frac{n_{lk}n_t}{n_{tk}n_l}\right\}+n_{rk}\,log\left\{\frac{n_{rk}n_t}{n_{tk}n_r}\right\}\right], \qquad (18)$$

where *l* represents node *t*'s data assigned to the left child node and *r*, the data assigned to the right child node. Note that deviance is zero when a node contains observations from only one response group.

Trees resulting from recursive partitioning are generally too large and tend to overfit the data. To reduce tree size, the effect of removing different terminal nodes (i.e., pruning the tree) on *tree* deviance, which is the sum of the deviance at each terminal node, is recursively evaluated. The routine stops pruning when the tree reaches the specified maximum size. This tree will have the lowest deviance of any tree of its size (Chou et al., 1989). To improve the predictive ability of tree models (i.e., reduce overfitting), optimum tree sizes can be determined by examining plots of the $\hat{E}ER$ by tree size (Brieman et al., 1984). These plots generally show an initially rapid decrease in error rate with increasing tree size, followed by relatively stable error rates, and then gradual increases in error as the larger trees begin overfitting the data. The most parsimonious tree model is generally considered the one in which size and expected error are minimized.

2.4. Nearest Neighbor Discriminant Analysis

2.4.1 Theory and definitions

K-Nearest Neighbor (KNN) classification is used to predict the response at a point $\mathbf{x}_o$ in the covariate space by first finding the nearest *K* observations to $\mathbf{x}_o$ and fitting a local response distribution. The response is then predicted with the mode of this local distribution (Hand 1989). KNN classification is relatively flexible and does not require an assumption of multivariate normality or strong assumption implicit in specifying a link

function (e.g., the logit link). It is based on the assumption that the characteristics of members of the same class should be similar and thus, observations located close together in covariate (statistical) space are members of the same class (e.g., Fig. 2) or at least have the same posterior distributions on their respective classes (Cover and Hart 1967). KNN has been compared against several neural network and nonparametric statistical methods and has been found to be among the better classifiers (Ripley 1993). One drawback however, is that KNN classification rules are difficult to interpret because they are only based on the identity of the *K* nearest neighbors. Therefore, information for the remaining *n* - *K* classifications is ignored (Cover and Hart 1967).

An extension of a nonparametric categorical regression smoother by Tutz (1990), referred to here as the extended K-nearest neighbor classifier (EKNN), is as follows. Let *Y* be the discrete dependent variable having *m* categories and let $\mathbf{x} = (z_1, z_2,..z_q, w_1, w_2,..w_r)'$ be the vector of covariates consisting of *q* quantitative covariates and *r* qualitative covariates. At a prediction location $\mathbf{x}_o$, estimate the response probability for category *j* with $\hat{\pi}_{0j} = \left(\sum_{i=1}^{K} y_{ij}\right) / K$, i.e., the relative local frequency of the j^{th} response.

By assuming zero correlation between the quantitative and qualitative covariates and themselves, define a distance measure between x_o and x_i as follows. First, define a vector of generalized differences: $\mathbf{s} \equiv D^{-1/2}\left(\mathbf{x}_0 - \mathbf{x}_i\right)$, where

$$D_{ii} = \begin{cases} \sqrt{Var\left[z_i\right]}, & i \le q \\ 1, & i > q \end{cases} \tag{19}$$

and

$$\mathbf{s} = \left(\mathbf{x}_0 - \mathbf{x}_i\right) \equiv \begin{bmatrix} \left|z_{01} - z_{i1}\right| \\ \vdots \\ \left|z_{0q} - z_{iq}\right| \\ d_w\left(w_{01}, w_{i1}\right) \\ \vdots \\ d_w\left(w_{0r}, w_{ir}\right) \end{bmatrix}. \tag{20}$$

The distance between qualitative covariates, which are assumed to be uncorrelated among themselves and with the quantitative covariates, is defined following Tutz (1990) as

$$d_w(w_{0j}, w_{ij}) \equiv \begin{cases} 0, w_{0j} = w_{ij} \\ 1, w_{0j} \neq w_{ij} \end{cases} . \quad (21)$$

Let V be the correlation matrix of the covariates:

$$V \equiv D^{-1/2} \begin{bmatrix} C_{qq} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} D^{-1/2}, \quad (22)$$

where C_{qq} is the within-category pooled variance-covariance matrix of the quantitative covariates. Then $d(\mathbf{x}_0, \mathbf{x}_i) = \sqrt{\mathbf{s}' V^{-1} \mathbf{s}}$ is the generalized Mahalanobis distance between $\mathbf{x}_0$ and $\mathbf{x}_i$ (Johnson and Wichern 1992). Note that if the assumption of independence among the qualitative covariates is not met, this distance computation will distort the true statistical distance between two observation locations (i.e., covariate vectors).

2.4.2 EKNN model fitting

The optimal value of *K* is defined as the value that minimizes the $\hat{E}ER$ and is found by repeating the cross-validation analysis for increasing values of *K*. Note that if the number of neighbors (*K*) is small, the estimated local response distribution will be based on a small sample and hence will have large variance. In the limit with *K* = 1, the estimated distribution will be degenerate at the response category of the nearest neighbor. If a lower variance estimate of the true response distribution at a particular location in the covariate space is desired, *m* delete-d jackknife samples (see below) may be drawn from the full sample and EKNN predictions of the response at the desired location computed from each jackknife sample. The percentage of predictions in each response category is the estimate of the response distribution at that location. Thus as *m* increases, the variance of each response category probability estimate decreases. Cover and Hart (1967) also show that for *K* = 1, the error rate of KNN approaches that of the optimal Bayes rule classifier as the sample size grows to infinity (see also Ripley, 1996).

Because EKNN classification is computationally fast, it lends itself to computer intensive hypothesis testing (see below). Further, unlike logistic regression, such testing can always be performed under a full model that consists of any number of covariates. That is, the use of all possible subsets model selection does not depend on the

computational tractability of a likelihood function. Additionally, missing values on the qualitative variables can be accommodated by either eliminating those observations with missing covariate values or by defining a "missing" category for each covariate that has missing values.

2.5 Modular Neural Networks

2.5.1 Theory and definitions

A neural network (NN) is a particular family of nonlinear functions, $g(\mathbf{x})$ that map the input vector of covariates, $\mathbf{x}$ to a k-valued response or *output*. Unlike parametric statistical models, most NN's do not explicitly model the error term and do not model the response variable's probability distribution. NN's are quite often extremely accurate (Cheng and Titterington, 1994; Anand et al., 1995), but difficult to interpret because the complex nature of their interconnected functions. They generally consist of four linked components: the input, hidden, and output layers, and the target (Fig. 3). The input layer is made up of covariate nodes (one for each) and a bias node used during neural network training. The hidden layer is composed of hidden nodes, each containing a set of link weights (one for each covariate and the bias term) that are analogous to parameter estimates in a generalized linear model (Sarle, 1994).

To illustrate the relationships between layers, consider the prediction of a response with the classical NN given an input vector $\mathbf{x}$. Predicted responses, are estimated using activation functions in both the hidden and output layers as follows. First, fix the values x_{p+1} and y_{L+1} to 1. The hidden layer output vector $\boldsymbol{y}_l$ is estimated using (say) a sigmoidal mashing function (i.e., logistic function bounded by 0-1) as:

$$y_l = \frac{exp(\mathbf{x}'\boldsymbol{\omega}_l)}{1+exp(\mathbf{x}'\boldsymbol{\omega}_l)},\ l = 1,..., L \tag{23}$$

where $\mathbf{x} = (x_1, ...,x_p, x_{p+1})'$ are the covariates and $\boldsymbol{\omega}_l = (w_{l,1}, ...,w_{l,p+1})'$ are the weights. Note that the $\boldsymbol{\omega}_{L+1}$ is the hidden layer bias, which is similar to a constant in a generalized linear model (Sarle, 1994). The output vector, $\boldsymbol{y}$, is then passed to the output layer and used to compute the output layer nodes values as:

$$z^*_j = \frac{exp(\mathbf{y}'\boldsymbol{\nu}_j)}{1+exp(\mathbf{y}'\boldsymbol{\nu}_j)}, \tag{24}$$

where $\boldsymbol{\nu} = (\nu_{j,1}, ..., \nu_{j,L}, \nu_{j,L+1})'$ are the link weights and z^*_j is the output value for module $j = 1,.., J$. The values of z^* are used to predict an observation's response, which is identified as the response with the largest z^*.

A modular neural network (MNN) differs from a NN in that each category has its own hidden layer module (e.g., Fig 3) which, through training, becomes "specialized" at predicting the associated category. Because of these hidden layer modules, a MNN has many more $\boldsymbol{\omega}$ vectors than a NN. Experiments reported in Anand et al. (1995) suggest that a MNN is often a more accurate classifier than a classical NN.

2.5.2 Constructing and training a MNN

The parameters of a MNN are L, $\boldsymbol{\omega}_1,..., \boldsymbol{\omega}_L$, and $\boldsymbol{\nu}_1,..., \boldsymbol{\nu}_k$. Hence, there are $1 + L(p + 1) + k(L + 1)$ parameters to be estimated from a sample. Because a NN can "grow" parameters quickly as p, L, or k become large, NN's are prone to data overfitting, i.e., producing a model that attempts to represent part of the noise in the data. On the other hand, the reason NN's are attractive is because for L large enough, a NN can be found that is equivalent to any function $g(\mathbf{x})$. This means that, given a large enough L, any relationship between $\mathbf{x}$ and $\mathbf{z}$ can theoretically be modeled with a NN (Hornik, 1991). This makes NN's attractive when $g(\mathbf{x})$ is suspected of being nonlinear with nonmultiplicative interactions. Thus when fitting a NN, one must contend with 2 basic problems: (1) how to efficiently estimate the parameters of the NN so as to minimize EER and (2) how to choose the optimal the value of L, so that overfitting does not occur but EER is minimized.

The first problem is called "neural network training" and is similar to statistical parameter estimation. Note that the statistical technique of parametric maximum likelihood estimation cannot be directly used because a NN is not a probability model of the observed process. Because of the typically large parameter count, efficient mathematical programming algorithms are needed to search for parameter values that minimize some measure of NN agreement with a data set. For the current study, we used quasi-Newton methods (Setiono and Hui, 1995) for the training step over the less-efficient Back-Propagation method. Although this routine is relatively fast and efficient, it can converge to a local minimum where classification accuracy is very low (Setiono

and Hui, 1995). We found that artificially setting one observation in the data set to "missing" (only) during the initial training helps to break free of potential local minima.

MNN training begins with 2 hidden nodes (L) per module (response category). Initial hidden node weights are randomly assigned and the quasi-Newton routine searches for parameter values that minimize the disagreement between the MNN and the data set as measured by:

$$S \equiv \sum_{i=1}^{k} \left[\frac{\sum_{j=1}^{n_i} s_j}{n_i} \right], \tag{25}$$

where n_i is the number of observations on the i^{th} category and, for the j^{th} observation ($\mathbf{x}_j$, $\mathbf{z}_j$):

$$s_j = \begin{cases} \left[1 - \max_j \left(\hat{z}_j^*\right)\right]^2, \hat{z}_j = z_j \text{ correct prediction} \\ \left[\max_j \left(\hat{z}_j^*\right)\right]^2, \text{otherwise} \end{cases} \tag{26}$$

Advantages of this measure of disagreement over (say) the energy measure used in Back-Propagation are that S is not overly-sensitive to radically different values of n_i since only error rates are summed and the ultimate, pragmatic measure of classifier goodness, the misclassification rate, is directly minimized to fit the MNN to the data set.

The second problem is called "network construction" and has been attacked with a variety of techniques (e.g., Setiono and Hui, 1995). Here, additional hidden nodes (L) are added in a stepwise manner and the MNN is retrained to increase its predictive ability. Thus, constructing an optimal (best predicting) MNN can be conducted in a manner similar to the selection of the optimal K for the EKNN models, with the optimal L considered to be the one in which the $\hat{\text{E}}\text{ER}$ is minimized.

2.6. Hypothesis testing with nonparametric classifiers

2.6.1 Theory and definitions

Classically, a hypothesis test is performed on the hypothesis that the effect of a set of excluded covariates is exactly zero. Such a test can be constructed with a computer intensive approach called a *Monte Carlo hypothesis test*. For a nonparametric classifier such as EKNN or a tree, such a test can be based on resampling statistics. See Hall and

Titterington (1989) for a description and asymptotic properties of Monte Carlo hypothesis tests, and Shao and Tu (1995) for a like discussion of resampling statistics.

Let the null hypothesis, H_0 be that there is no difference in the error rate between a reduced model and the full model. Let $\delta \equiv \text{EER}_R - \text{EER}_F$ where EER_R and EER_F are the expected error rates using the reduced set of covariates and the full set of covariates, respectively. Define the test statistic to be $T \equiv \hat{\delta} - \delta$. Note that under H_0, $T = \hat{\delta}$.

We used EER as the basis for the test statistic because we believe that for purposes of both understanding ecological processes and developing natural resource management strategies, the most important property of a data-based classifier is its out-of-sample prediction accuracy. That is, if performance on this criterion is not acceptable, the classifier should not be used for policy-relevant decision-making.

If the number of observations on each response category are very different, a possible drawback to using EER as the test statistic is that the test will only be sensitive to error rate changes in the most frequent category or categories. This may result in declaring as insignificant covariates that in actuality, significantly affect the error rate of the more rare categories. Therefore, a second test statistic, which we used for all our analyses, is defined based on $\text{EERS} \equiv \sum_{i=1}^{k} \text{EER}_i$ yielding $\delta_S = \text{EERS}_R - \text{EERS}_F$ and $T_S = \hat{\delta}_s - \delta_s$. The T_S statistic is uniformly sensitive to changes across individual category error rates no matter what the observed frequencies are and hence, can be said to be less sample-dependent.

The Monte Carlo hypothesis test procedure is as follows.

Step 1: Compute $\hat{EER}_R$ and $\hat{EER}_F$ from the actual data set (hereafter called the full sample). Compute $T = \hat{\delta}$, the observed value of the test statistic assuming H_0 is true.

Step 2: Sample without replacement r ($< n$) observations from the full sample.

Step 3: Compute reduced and full model error rate estimates of the classifier using this jackknife sample. Denote these two error rate estimates EER_R^* and EER_F^* respectively. Compute and store $T^* = \hat{\delta}^* - \hat{\delta}$, the jackknife sample's test statistic value. Note that the true (but unknown) error rates have been

replaced with those estimated from the full sample. Doing so gives the Monte Carlo test good statistical power (Hall and Titterington 1989).

Step 4: Repeat steps 2 and 3 m times (always with a new randomly selected jackknife sample).

Step 5: Compute the p-value of the test to be the fraction of T^* values greater than T.

Note that when $r < n - 1$, the histogram of the m T^* values is a *delete-d jackknife* statistic (Shao and Tu 1995) where $d = n - r$. For properly chosen values of d and m, delete-d jackknife sampling can approximate the true test statistic distribution (Shao and Tu, 1995). Additionally, although the cross-validation $\hat{EER}$ is nearly unbiased, defining the test statistic as a function of only *differences* in estimated error rates reduces the effect on the test statistic of any such biases.

The empirical distribution of T, needed for Step 5, above, is formed with delete-d jackknife samples instead of bootstrap samples (i.e., sampling with replacement) because numerical experiments showed such samples produced excessively biased values of $EER_{(.)}^*$ (also see Hall and Titterington 1989). In addition, subsamples for computing the test statistic's empirical distribution are formed from resampling instead of simulation because with $K = 1$, EKNN's locally estimated response distributions are all degenerate at the nearest neighbor's response value and hence all simulated data sets from these response distributions would be identical.

2.6.2 Evaluation of Monte Carlo hypothesis test consistency

For the hypothesis test to be consistent, both d and m need to be large (Shao and Tu, 1995). That is, jackknife sample sizes (r) should be small relative to the total sample size (n) and the number of jackknife samples (m) should be large. Total sample sizes, however, are likely to vary considerably among datasets for most practical applications. Consequently, the use of a single r for all analyses could result in inconsistent or unreliable hypothesis tests, whereas large m could result in excessively long computer run-times. We examined the relationships between hypothesis test p-values and r and m for 3 salmonid taxa representing small (Yellowstone cutthroat trout), moderate (redband trout), and large (steelhead) sample sizes. Using the EKNN models and a randomly chosen covariate for each species, the influence of r on p-values was examined for each

of the 3 taxa by initially setting $r = 0.05n$ and running 100-replicate tests with $m = 100$. This process was repeated with r increasing in $0.05n$ increments until it reached $0.5n$. The influence of m on the variability of p-values was examined by running 100-replicate tests for m from 50-100 with $r = 0.15n$.

The influence of r on test p-values was similar among species with relatively stable p-values for r from 10-20% of the total sample sizes (Fig. 4a). Similarly, p-values were most variable for all salmonid tests with 50 jackknife samples (Fig. 4b). The relatively small p-values for the steelhead tests, with a mean of 0.02, were also the most variable but tended to stabilize (shallower slope) after approximately 500-600 samples. Based on these results, we used $r = 0.15n$ and $m = 500$ for all of the Monte Carlo hypothesis tests with the salmonid response data below.

3. Model fitting and parameter estimation

3.1. Generalized Logit Model

Preliminary modeling indicated that the cumulative logit model was inappropriate for modeling population status of all salmonid taxa (i.e., all models failed the parallel lines test). Hence, we restrict our examination of logit model performance to the multinomial logit model (1). All of the modeling (below) was conducted using CATDAT statistical software (Haas et al., 1999) available via the internet (freeware).

Following a recommendation in Agresti (1990), the most frequent response for each data set was defined to be the J^{th} response (i.e., baseline) category. Levels of the qualitative covariate, *Mgntcls* (Table 1), were recoded into dummy variables (0,1) prior to fitting each generalized logit model. For most species, some *Mgntcls* levels were rare (i.e., comprised < 10% of observations), which could have caused unstable maximum likelihood estimates (Agresti 1990). Consequently, we combined the observations from rare, but related, land management types into composite groups to increase the number of observations and maintain the interpretability of model coefficients (Table 2). For instance, data from relatively unimpacted areas, National Park (*Np*) and Forest Service wilderness (*Fw*) lands, were combined into a single group *Np-Fw*.

The choice of covariates and the form of the generalized logit model can influence model performance (i.e., out-of-sample prediction accuracy). For example, including too few covariates in a model can result in an incomplete representation of the

factors that affect the responses, whereas including too many or insignificant covariates could introduce too much noise in the model, lowering performance. Similarly, interactions among the covariates may also be important for characterizing responses. To obtain the best generalized logit model for each taxon, we fit 4 models: (1) full main-effects, (2) statistically significant main-effects, and (3) full main-effects and statistically significant 2-way interactions, and (4) statistically significant main-effects and 2-way interactions; and examined their expected error rates. Statistically significant main-effects (only) were selected via backward elimination, whereas forward stepwise selection was used to select statistically significant main effects and 2-way interactions. To maintain a consistent 0.05 experiment-wise error rate, covariates and interactions were considered statistically significant at a Bonferroni adjusted $\alpha = 0.05/k$ (k = the number of main effects and/or interactions). The overall statistical significance of each model was assessed with log-likelihood test statistic and the goodness-of-fit of each model was assessed by examining the studentized Pearson residuals and via the Osius and Rojek (1992) increasing cell asymptotics and Andrews' (1988) omnibus chi-square tests.

3.2. Classification Tree

For each salmonid taxon, optimal tree sizes were determined by fitting models with all covariates (Table 1) and setting minimum partition sizes (stopping value) at $0.04n$. EERs were then estimated for trees ranging in size from 10-70 nodes or the maximum number of nodes possible for smaller data sets (e.g., 40 nodes for ocean-type chinook salmon). Tree sizes that resulted in the lowest overall $\hat{E}ER$ (e.g., Fig. 5a) were considered optimal and were used for the Monte Carlo hypothesis testing procedure to ensure that the expected resolution of the subsampled trees were comparable to the full trees. Statistically significant covariates ($P<0.05$) were selected via backward elimination with the Monte Carlo hypothesis testing procedure described above.

3.3 Extended K-nearest neighbor models

Extended *K*-nearest neighbor models were initially fit for each salmonid using all covariates (Table 1). The optimal numbers of neighbors (K) were determined by fitting models with K set from 1- 30 and examining overall $\hat{E}ER$ (e.g., Fig. 5b). Similar to the tree models, the optimal value was considered to be that which had the lowest

overall $\hat{E}ER$. These were then used for the Monte Carlo hypothesis testing procedure for selecting statistically significant covariates ($P<0.05$) via backward elimination.

3.4 Modular neural network models

The modular neural networks (MNN) were fit for each salmonid using all covariates (Table 1). The optimal number of hidden nodes (L) was determined by fitting models with the number of nodes set from 1-15 and estimating the overall $\hat{E}ER$ (e.g., Fig. 5c). Similar to the other nonparametric classifiers, the optimal L was considered to be that which had the lowest overall $\hat{E}ER$. Fitting and cross-validation of the MNN required excessively long run times. For example, leave-one-out cross-validation for the steelhead 15 node MNN required approximately 62 hours on a *RISC System 6000 Uni-processor*. Therefore, we did not conduct Monte Carlo hypothesis tests with the MNN models.

4. Results

4.1 Generalized logit model

For all salmonid taxa, each of the 4 generalized logit models was statistically significant (log-likelihood test statistic, $P < 0.05$). However, the full main-effects and statistically significant 2-way interaction models consistently had the lowest overall EER_w, across species (Table 3). Using the best fitting models for each taxon (except Yellowstone cutthroat trout), we found several of the estimated probabilities (>23 per taxon) for *strong* and *depressed* responses were very small ($<10^{-5}$) and asymptotic variances were very high ($\sigma_1^2 > 10^{13}$). Hence, condition 1 of the increasing-cells asymptotics did not hold and the power of the Osius and Rojek goodness-of-fit test was unacceptably low. The Andrews omnibus χ^2 test, however, indicated that the logit models failed to fit the trinary response data for all of the resident salmonids ($P< 0.05$), except Yellowstone cutthroat trout ($P= 0.63$), and the quaternary response models for all of the anadromous salmonids ($P<0.05$). Because the very small probabilities were for the less-frequent *strong* and *depressed* responses, the failed increasing-cells asymptotics might have been due to these categories. Thus, we combined the *strong* and *depressed* population status into a single category, *present*, and refit the best model for each taxon except Yellowstone cutthroat trout. Andrews omnibus χ^2 test of these models indicated a

better fit (P> 0.05) for the binary response data of the resident salmonids and the trinary response data of the anadromous salmonids.

Although the resident and anadromous salmonid binary and trinary response data, respectively, could be fit by a generalized logit model, leave-one-out cross-validation indicated that the logit models were poor at predicting salmonid population status (Table 4). Among the salmonids, the logit model was best at predicting the population status of stream-type chinook salmon (39.0% $\hat{E}ER$) and worst at predicting steelhead and Yellowstone cutthroat trout status (56.7 and 56.8% $\hat{E}ER$, respectively). Category-wise error rates also indicated that, in most instances, the logit model predictions were unreliable for the population responses with the fewest observations. For instance, migrant status for the anadromous salmonids had the fewest observations and the greatest classification and prediction error rates for all of the logit models (Table 4). A similar pattern was apparent for the resident salmonid binary response models.

4.2. Classification tree

Optimal tree sizes varied considerably among species and appeared to be unrelated to $\hat{E}ER$ or the number of response categories (Table 4). For example, ocean-type and stream-type chinook salmon models had lowest overall $\hat{E}ER$ and smallest and largest optimal tree sizes, respectively, whereas the redband trout model had second smallest optimal tree size and the greatest overall $\hat{E}ER$.

Following model selection, an examination of the $\hat{E}ER$ indicated an improvement in classification accuracy for 5 of the 7 tree models (Table 6). Overall, the Yellowstone cutthroat trout model had the greatest increase in accuracy with a 23.3% reduction in the $\hat{E}ER$ followed by stream-type chinook salmon (8.4%) and bull trout (7.5%). The classification tree models had, on average, fewer statistically significant covariates than the generalized logit models (Table 6). There was also considerable overlap in the identity of the significant covariates among taxon-specific models fit with the 2 classifiers. For example, all of the statistically significant covariates for bull trout, Yellowstone cutthroat trout and the anadromous salmonid classification tree models were

also found to be significant for the generalized logit models (Table 5). The overall and category-wise $\hat{E}ER$ for the tree models were also, on average, 10-20% lower than those of the generalized logit models (Table 4). However, response categories with the fewest observations tended to have the highest category-wise error rates, which was similar to the logit models.

Although classification tree size appeared unrelated to accuracy, smaller trees were generally easier to summarize and interpret. For example, the landscape characteristics associated with ocean-type chinook salmon absence could be summarized with 3 rule sets (Fig. 6): (1) elevation greater than 2075 m and fewer than 1823 contributing upstream subwatersheds; and (2) elevation less than 2076 m and between 264 and 1051 contributing upstream sub-watersheds or (3) fewer than 10 contributing upstream subwatersheds and annual precipitation less than 234 mm. Thus, the tree model suggests that ocean-type chinook salmon are generally absent (n=298) in higher elevation subwatersheds containing smaller streams (i.e., fewer contributing upstream subwatersheds) and in lower evaluation subwatersheds with containing only small to moderately sized streams. In contrast, summarizing the landscape characteristics associated with Yellowstone cutthroat trout absence with the larger model would require 6 rule sets and would be likewise more difficult to interpret (Fig. 7).

4.3 Extended K-nearest neighbor models

Optimal K for the salmonid models were relatively low and varied from among species (Table 4) and, similar to the classification tree best parameter, the values of the optimal K appeared to be unrelated to the $\hat{E}ERs$ or the number of response categories. The Monte Carlo hypothesis test of covariates indicated that the EKNN models also had, on average, fewer statistically significant covariates than the generalized logit models (Table 6), and there was considerable overlap in the identity of the significant covariates among the all three classifiers. The $\hat{E}ERs$ however, suggested that the EKNN models were more accurate than both the classification tree (except ocean-type chinook salmon) and the generalized logit models (Table 4). On average, the overall $\hat{E}ER$ of the EKNN models were 56 % and 8% lower than the logit and tree models, respectively, but the responses with the fewest observations still tended to have the highest category-wise

classification and prediction $\hat{EER}$. Among species, the EKNN models were most accurate at predicting the population status for all anadromous salmonids and poorest at predicting redband and westslope cutthroat trout status (Table 4).

Although relationships between the covariates and the responses could not be determined with the EKNN models, the relationships (i.e., similarities) among responses were examined using the mean Mahalanobis distances (Table 7). Among resident salmonids, the landscape characteristics of subwatersheds with no populations (absent) generally differed most from those with strong and depressed populations. Differences among landscape characteristics associated with anadromous salmonid population status, however, were greatest between the subwatershed containing only migratory corridors (migrant). Additionally, the average Mahalanobis distances also appeared unrelated to the accuracy of the EKNN models (Tables 4 and 7).

4.4 Modular neural network models

The optimal number of hidden nodes (L) varied little among salmonid taxa and ranged from 7-11 (Table 4). The MNN had the lowest overall and category-wise $\hat{EER}$ of any of the classifiers considered in this analysis. On average, the overall EERs of the MNN models were 69%, 45%, and 32% lower than the logit, tree, and EKNN models, respectively. Among species, the MNN models were most accurate at predicting the population status of ocean-type chinook and Yellowstone cutthroat trout and poorest at predicting redband and westslope cutthroat trout. In addition, the category-wise error rates were unrelated to the number of observations, which was in sharp contrast to the other classification techniques (Table 4).

5. Discussion

An attractive feature of a parametric approach is that hypothesis testing is theoretically developed and computationally convenient. The usefulness of results from such tests however, is compromised if the parametric model's fit to the data set is poor. Indeed, the generalized logit model was by far the poorest performing technique considered (Table 4). Further, the salmonid population response processes did not appear to satisfy either increasing-cell-counts nor increasing-cells asymptotics assumptions. Had these conditions not been examined, we could have falsely concluded that the original

(uncombined) responses could be fit by a generalized logit model based on the Osius and Rojek tests. This highlights the need to verify that all conditions used to establish the asymptotic distribution of a test statistic are been met by the observed process. Such conditions may not always be mentioned in a brief description of the test, say in the documentation of a statistical software package. Instead, the delineation of all such conditions may require a close reading of the test's statistical derivation.

Reasons for the poor performance of the logit models were likely due to the presence of nonlinear relationships and nonmultiplicative interactions, effects that nonparametric classifiers are potentially able to capture (Chambers and Hastie, 1992). The much greater accuracy of the nonparametric methods, particularly the MNNs (Table 4), tends to support this contention. MNNs are universal approximators (White 1992) and thus, were likely able to capture complex relationships between the landscape covariates and salmonid population responses. MNNs also tended to have uniformly low error rates across response categories for all taxa, whereas category-wise error rates for the other 3 methods were inversely related to category-specific sample sizes (Table 4). Unlike the other methods, MNNs model each response separately and hence, were able to specialize in predicting individual responses. The substantial differences in accuracy of the various methods highlights the fact that model accuracy can be significantly influenced by how well a particular technique approximates the biological response of interest. Thus, researchers should exercise caution when attempting to interpret poorly fitting models in terms of weak or non-existent biotic responses.

Although the MNNs were the most accurate method considered, they were essentially useless for examining the relationships between the landscape covariates and salmonid population responses. They also required extensive computer resources for model fitting and cross-validation, which prevented us from using the Monte Carlo hypothesis test to gain some insight into these relationships. MNNs also do not generate probabilistic estimates for each response, which is unlike the other methods considered. Thus, they are inappropriate for use in situations that require an explicit expression of uncertainty, such as risk assessments. However when very accurate estimates are required, we believe that the MNN is an ideal method.

Trees offer more interpretability than MNNs (Figs. 6 and 7) and allow more complex interactions to be captured, but they can exhibit higher error rates due to their forward-selection mode of construction. Some of this error can be reduced if only the significant covariates are used to fit the tree (Table 5), possibly due to the reduction in aliasing. In contrast, the EKNN classifiers were relatively accurate (Table 6), but they lack interpretability. As we have shown, the Monte Carlo hypothesis tests can be used to provide some insight into the significance of individual covariates with the EKNN. However, they cannot be used to determine form and strength of the relationships. Additional insight also can be gained by examining the relationships among responses via the mean Mahalanobis distance, which is similar to other distance (similarity) measures For example, relationships among several responses can be investigated with hierarchical cluster analysis and multidimensional scaling. In addition, the relative speed at which EKNN models can be fit (compared to the other nonparametric methods) in combination with their relatively high accuracy suggests that it is an ideal method for preliminary analyses.

As we have shown, a parametric model is no longer a prerequisite to formal, statistical hypothesis testing. These hypothesis tests also provide additional insight and can improve the performance of the nonparametric approaches. Across taxa, all of the statistically significant covariates for each method were biologically plausible (see Lee at al., 1997 for a review). For instance, bull trout are negatively affected by anthropogenic impacts (Rieman et al., 1997; Baxter et al., 1999 and references therein) and both measures of human impacts, road density and dominant land management type, were found to be significant for all methods considered (Table 4). Furthermore, the large number of significant covariates common among methods suggests that each classifier is capable of detecting similar relationships (patterns) in the data. Thus, greater ecological insight could be gained by using several different classification methods.

Conclusions

Classification accuracy and model interpretability are among the most desirable characteristics of statistical classification methods. Of the techniques considered, the MNN and EKNN were the most accurate classifiers (Table 4) but were uninterruptible, whereas the generalized logit models were readily interpretable but were inaccurate. In

contrast, the tree models were fairly accurate (Table 4) and interpretable (Figs. 6 and 7) and would be the best method for modeling salmonid population status if both properties were required from a single model. The substantial overlap among the significant covariates for each classifier (Table 6) also suggests that each is capable of detecting similar relationships between responses and covariates. Thus if a single model is not required, an alternative approach could be to use two or more models. For example, accurate predictive models could be created with the EKNN or MNN and the significance of individual covariates could be examined via the Monte Carlo hypothesis test. The relationship between the significant covariates and the responses could then be examined with the generalized logit model and/or the classification tree. Consequently, employing several classifiers may provide greater biological insight without sacrificing prediction accuracy.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Anand, R., Mehrotra, K., Mohan, C. K., and Ranka, S., 1995. Efficient classification for multiclass problems using modular neural networks. IEEE Transactions on Neural Networks, 6l: 117-124.

Andrews, D. W. K., 1988. Chi-square diagnostic tests for econometric models. Journal of Econometrics 37: 135-156.

Baxter, C.V., Frissell, C. A., and F. R. Huaer, 1999. Geomorphology, logging roads, and the distribution of bull trout spawning in a forested river basin: implications for management and conservation. Transactions of the American Fisheries Society 128: 854-867.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J., 1984. Classification and Regression Trees. Wadsworth International, Belmont, CA.

Carroll, C., Zielinski, W. J., and Noss, R. F., 1999. Using presence-absence data to build and test spatial habitat models for the fisher in the Klamath Region, U.S.A. Conservation Biology 13: 1344-1359.

Chambers, J. M. and Hastie, T. J., 1992. Statistical Models in *S*. Wadsworth and Brooks, Pacific Grove, CA.

Cheng, B. and Titterington, D. M., 1994. Neural networks: a review from a statistical perspective. Statistical Science 9l: 2-54.

Chou, P. A., Lookabaugh, T., and Gray, R. M., 1989. Optimal pruning with applications to tree-structured source coding and modeling. IEEE Transactions on Information Theory 352: 299-310.

Christensen, R., 1990. Log-Linear Models. Springer-Verlag, New York.

Cochran, W. G., 1954. Some methods of strengthening the common χ^2 tests. Biometrica 10: 417-451.

Cover, T. M. and Hart, P. E., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13: 21-27.

Dunham, J. B. and Rieman, B. E., 1999. Metapopulation structure of bull trout: influences of habitat size, isolation, and human disturbance. Ecological Applications 9:642-655.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 783: 316-331.

Fahrmeir, L., 1988. A note on asymptotic testing theory for nonhomogeneous observations. Stochastic Processes and Their Applications 28: 267-273.

Fahrmeir, L. and Kaufmann, H., 1987. Regression models for non-stationary categorical time series. Journal of Time Series Analysis, 82: 147-160.

Fukunaga, K., and Kessell, D., 1971. Estimation of classification error. IEEE Transactions on Computers C-20: 1521-1527.

Fahrmeir, L. and Tutz, C., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag, New York.

Gnanadesikan, R., Blashfield, R. K., Brelman, L., Dunn, 0. J., Friedman, J. H., Fu, K-S, Hartigan, J. A., Kettenring, J. R., Lachenbruch, P. A., Olshen, R. A., and Rohlf, F. J., 1989. Discriminant analysis and clustering. Statistical Science 4l: 34-69.

Haas, T.C., Peterson, J.T., and Lee, D.C., 1999. CATDAT- a program for parametric and nonparametric categorical data analysis. <http://www.fs.fed.us/rm/boise/fish/catdat/catdat.html>

Haberman, S. J., 1988. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. Journal of the American Statistical Association 402: 555-560.

Hall, P. and Titterington, D. M., 1989. The effect of simulation order on level accuracy and power of Monte Carlo tests. Journal of the Royal Statistical Society, Series B 48: 459-467.

Hand, D.J., 1882. Kernel Discriminant Analysis. Research Studies Press, New York.

Healey, M. C., 1991. Life history of chinook salmon (*Oncorhynchus tshawytscha*). In: Groot, C. and Margolis, L. (Eds), Pacific salmon life histories. University of British Columbia Press, Vancouver, pp. 311- 393.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neutral Networks 4: 251-257.

Hosmer, D. W. and Lemeshow, S., 1980. Goodness of fit tests for the multiple logistic regression models. Communications in Statistics - Theory and Methods, 910: 1043-1069.

Johnson, R. A. and Wichern, D. W., 1992. Applied Multivariate Statistical Analysis, 3rd edition. Prentice-Hall, Englewood Cliffs, NJ.

Kruse, C. G., Hubert, W. A., and Rahel, F. J., 1997. Geomorphic influences on the distribution of Yellowstone cutthroat trout in the Absaroka Mountains, Wyoming. Transactions of the American Fisheries Society 126:418-427.

Lachenbruch, P. A., 1965. Estimation of error rates in discriminant analysis. Ph.D. dissertation, University of California at Los Angeles.

Lachenbruch, P. A., 1975. Discriminant Analysis. Hafner, Press, Macmillan, New York.

Lee, D.C., Sedell, J.R., Rieman, B.E., Thurow, R. F., Williams, J.E., 1997. Broadscale assessment of aquatic species and habitats. An assessment of ecosystem components in the interior Columbia Basin and portions of the Klamath and Great basins. U.S. Forest Service General Technical Report PNW-GTR-405. Pacific Northwest Research Station, Portland, OR.

Lek, S. and Guegan, J. F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120: 65-73

Levin, S. A., 1992. The problem of pattern and scale in ecology. Ecology 73: 1943-1967.

Miller, R. I., Stuart, S. N., and Howell, K. M., 1989. A methodology for analyzing rare species distribution patterns utilizing GIS technology: the rare birds of Tanzania. Landscape Ecology 2:173-189.

Naugle, D.E., Higgins, K. F., Nusser, S. M., and Johnson, W. C., 1999. Scale-dependent habitat use in three species of prairie wetland birds. Landscape Ecology 14: 267-276

Osius, G. and Rojek, D., 1992. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. Journal of the American Statistical Association 874:1145-1152.

Press, J. and Wilson. S., 1978. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 73:699-705.

Quigley, T.M. and Arbelbide, S.J. (Technical Editors), 1997. An assessment of ecosystem components in the Interior Columbia River Basin and portions of the Klamath and Great Basins. U.S. Forest Service, Pacific Northwest Research Station General Technical Report PNW-GTR-405, Portland, OR.

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

Ripley, B. D., 1993. Neural networks and flexible regression and discrimination. In: Mardia, K. V. and Kanji, G.K. (Eds.), Statistics and Images, Advances in Applied Statistics Series 1. Carfax, Abingdon.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, U.K.

Sarle, W. S., 1994. Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference. SAS Institute, Cary, NC.

SAS Institute, 1989. SAS/STAT User's Guide, Version 6, Fourth Edition, Vols. 1 and 2, SAS Institute, Cary, NC.

Setiono, R. and Hui, L. C. K., 1995. Use of a quasi-Newton method in a feedforward neural network construction algorithm. IEEE Transactions on Neural Networks 49: 273-277.

Shao, J. and Tu, D., 1995. The Jackknife and Bootstrap. Springer-Verlag, New York.

Thurow, Russell F., Lee, D. C., and Rieman, B. E., 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.

Tutz, G., 1990. Smoothed categorical regression based on direct kernel estimates. Journal of Statistical Computation and Simulation 36: 139-156.

White, H., 1992. Artificial Neural Networks: Approximation and Learning Theory. Blackwell, Oxford, U.K.

Table 1. Covariates used to parameterize the parametric and nonparametric models of population status for the 7 salmonid taxa.

Covariate Name	Description
Quantitative	
Bank	streambank erosion hazard
Baseero	base erosion index
Drnden	drainage density (km/km^2)
Elev	mean elevation (m)
Hk	soil texture coefficient
Hucorder	number of contributing upstream subwatersheds
Mtemp	mean annual temperature (^{o}C)
Pprecip	mean annual precipiation (mm)
Rdmean	mean road density (km/km^2)
Slope	area weighted average midslope (degrees)
Solar	(Langley's) mean annual solar radiation loading
Qualitative	
Mgntcls	dominant land management type, ten levels shown below
Br	Bureau of Land Management (BLM) rangeland
Fg	Forest Service (FS) forest and rangeland, moderate impact,
Fh	FS forest, high impact, grazed
Fm	FS forest, high-moderate impact, no grazing
Fw	FS managed wilderness
Np	National Park Service forest land
Pa	private agriculture
Pf	private land and FS forest land
Pr	private and BLM rangeland
Tl	tribal lands

Table 2. The dominant land-management types that were combined and dummy coded (0,1) for the generalized logit models of salmonid population status. Note that the last composite management type listed received a zero coding for all dummy variables.

Bull trout					
Fg-Fh	Fm	Fw-Np	Pa	Pf-Tl	Br-Pr
Redband trout					
Br	Fg-Fh	Pa	Pf-Tl-Fm	Pr	Fw-Np
Westslope cutthroat trout					
Fg-Fh	Fm	Fw-Np	Pf-Tl	Br-Pa-Pr	
Yellowstone cutthroat trout					
Fg-Fh	Fw-Np	Pa	Br-Fm-Pf-Pr-Tl		
Ocean-type chinook salmon					
Pa	Pf-Tl-Fm	Pr-Br	Fg-Fh-Fw-Np		
Stream-type chinook salmon					
Br	Fg-Fh	Pa	Pf-Tl-Fm	Pr	Fw-Np
Steelhead					
Br	Fg-Fh	Pa	Pf-Tl-Fm	Pr	Fw-Np

Table 3. Overall within-sample classification error rates (EER_w) for various generalized logit models of salmonid population status. Because of the negative bias of EER_w, cross-validation error rates can only be expected to be higher. Statistically significant covariates can be found in Table 4.

Logit model	Bull trout	Redband trout	Westslope cutthroat trout	Yellowstone cutthroat trout	Chinook salmon		Steelhead
					ocean-type	stream-type	
Main-effects	0.297	0.382	0.303	0.206	0.128	0.199	0.273
Statistically significant main effects	0.303	0.396	0.300	0.226	0.208	0.203	0.284
Main-effects and significant interactions	0.262	0.320	0.276	0.164	0.132	0.180	0.223
Statistically significant main-effects and interactions	0.300	0.354	0.283	0.218	0.143	0.223	0.230

Table 4. Summary of cross-validation classification and prediction (in parenthesis) error rates for each classifier and the optimal number tree nodes, nearest neighbors (K), and hidden nodes.

Resident species	Population status	N	Generalized logit model	Classification tree	*K*-nearest neighbor	Modular neural network
	overall			58 *nodes*	*K* = 7	8 *nodes*
Bull trout	EER		0.502	0.299	0.234	0.207
	Strong	169		0.674 (0.530)	0.544 (0.398)	0.112 (0.503)
	Depressed	624	0.779 (0.453)[1]	0.319 (0.501)	0.442 (0.373)	0.213 (0.343)
	Absent	1555	0.360 (0.382)	0.251 (0.155)	0.117 (0.175)	0.214 (0.059)
	overall			22 *nodes*	*K* = 9	9 *nodes*
Redband trout	EER		0.420	0.323	0.300	0.269
	Strong	250		0.692 (0.388)	0.668 (0.503)	0.188 (0.482)
	Depressed	712	0.332 (0.402) [1]	0.282 (0.393)	0.294 (0.342)	0.354 (0.263)
	Absent	825	0.524 (0.448)	0.246 (0.240)	0.194 (0.222)	0.219 (0.165)
Westslope	overall			52 *nodes*	*K* = 1	11 *nodes*
cutthroat trout	EER		0.469	0.281	0.245	0.220
	Strong	330		0.488 (0.329)	0.406 (0.382)	0.330 (0.338)
	Depressed	988	0.388 (0.223)	0.101 (0.272)	0.172 (0.185)	0.171 (0.179)
	Absent	271	0.860 (0.472) [1]	0.683 (0.295)	0.314 (0.306)	0.262 (0.298)
Yellowstone	overall			38 *nodes*	*K* = 3	10 *nodes*
cutthroat trout	EER		0.568	0.213	0.190	0.064
	Strong	173	0.594 (0.698)	0.179 (0.177)	0.178 (0.144)	0.050 (0.020)
	Depressed	115	0.809 (0.463)	0.384 (0.317)	0.244 (0.326)	0.061 (0.129)
	Absent	101	0.393 (0.505)	0.079 (0.174)	0.162 (0.110)	0.075 (0.042)
Anadromous species						
Ocean-type chinook	overall			21 *nodes*	*K* = 3	10 *nodes*
salmon	EER		0.402	0.101	0.170	0.021
	Strong	21		0.476 (0.389)	0.619 (0.579)	0.000 (0.125)
	Depressed	57	0.654 (0.775)[1]	0.386 (0.146)	0.491 (0.453)	0.018 (0.082)
	Migrant	59	0.983 (0.938)	0.029 (0.081)	0.074 (0.105)	0.000 (0.017)
	Absent	340	0.244 (0.246)	0.102 (0.102)	0.254 (0.170)	0.026 (0.003)
Stream-type chinook	overall			65 *nodes*	*K* = 3	8 *nodes*
salmon	EER		0.390	0.184	0.148	0.110
	Strong	8		1.000 (1.000)	0.625 (0.625)	0.000 (0.429)
	Depressed	470	0.872 (0.895)[1]	0.557 (0.377)	0.398 (0.390)	0.089 (0.397)
	Migrant	254	1.000 (1.000)	0.464 (0.343)	0.319 (0.244)	0.016 (0.269)
	Absent	2293	0.222 (0.264)	0.074 (0.145)	0.076 (0.089)	0.125 (0.008)
Steelhead	overall			62 *nodes*	*K* = 5	7 *nodes*
	EER		0.567	0.296	0.142	0.121
	Strong	23		1.000 (0.000)	0.870 (0.500)	0.696 (0.385)
	Depressed	969	0.678 (0.783)[1]	0.149 (0.475)	0.185 (0.222)	0.131 (0.193)
	Migrant	239	0.992 (0.959)	0.288 (0.260)	0.268 (0.236)	0.247 (0.227)
	Absent	1940	0.458 (0.448)	0.364 (0.099)	0.096 (0.087)	0.093 (0.069)

[1]Strong and depressed population status were combined in to the single response, p*resent*.

Table 5. Overall cross-validation error rates for full (all covariates) and reduced (significant covariates) classification tree models.

Resident salmonids	Full Model	Reduced Model
Bull trout	0.323	0.299
Redband trout	0.326	0.323
Westslope cutthroat trout	0.293	0.281
Yellowstone cutthroat trout	0.278	0.213
Anadromous salmonids		
Ocean-type chinook salmon	0.101	0.101
Stream-type chinook salmon	0.201	0.184
Steelhead	0.296	0.295

Table 6. Statistically significant covariates for the salmonid population status models, by taxa and method. A description of the covariates can be found in Table 1.

Resident salmonids	Generalized logit model	Classification tree	K-nearest neighbor
Bull trout	Slope, Solar, Rdmean, Mgntcls	Slope, Pprecip, Mtemp, Solar, Rdmean, Mgntcls	Slope, Drnden, Pprecip, Mtemp, Solar, Rdmean, Mgntcls
Redband trout	Slope, Drnden, Bank, Solar, Rdmean, Mgntcls	Hucorder, Slope, Bank, Baseero, Solar, Mgntcls	Elev, Slope, Bank, Solar, Rdmean, Mgntcls
Westslope cutthroat trout	Elev, Slope, Bank, Baseero, Hk, Mtemp, Solar, Mgntcls	Hucorder, Elev, Bank, Pprecip, Mtemp, Mgntcls	Hucorder, Bank, Baseero, Pprecip, Mtemp, Solar, Mgntcls
Yellowstone cutthroat trout	Hucorder, Elev, Slope, Baseero, Hk, Pprecip, Mtemp, Solar	Hucorder, Elev, Pprecip, Mtemp, Solar	Hucorder, Hk, Mtemp, Solar
Anadromous salmonids			
Ocean-type chinook salmon	Hucorder, Elev, Pprecip	Hucorder, Elev, Pprecip, Rdmean	Hucorder, Elev
Stream-type chinook salmon	Hucorder, Elev, Slope, Drnden, Bank, Pprecip, Mtemp, Solar, Rdmean, Mgntcls	Hucorder, Elev, Drnden, Bank, Pprecip, Solar, Mgntcls	Hucorder, Elev, Drnden, Bank, Mtemp, Solar, Mgntcls
Steelhead	Hucorder, Elev, Slope, Drnden, Bank, Baseero, Pprecip, Mtemp, Solar, Mgntcls	Hucorder, Elev, Slope, Baseero, Mntemp, Solar Mgntcls	Hucorder, Elev, Drnden, Bank, Pprecip, Mntemp

Table 7. Mean Mahalanobis distance between response categories from K-nearest neighbor classification of salmonid population status.

Resident salmonids		Strong	Depressed	
Bull trout	Depressed	2.8441	-	
	Absent	7.1477	4.5480	
Redband trout	Depressed	2.5263	-	
	Absent	6.4816	3.9974	
Westslope cutthroat trout	Depressed	3.7678	-	
	Absent	6.2722	2.8891	
Yellowstone cutthroat	Depressed	7.1240	-	
	Absent	14.5497	7.7357	
Anadromous salmonids		Strong	Depressed	Migrant
Ocean-type chinook salmon	Depressed	2.3306	-	
	Migrant	3.4599	1.9189	-
	Absent	1.3043	1.6020	3.0771
Stream-type chinook salmon	Depressed	3.4135	-	
	Migrant	8.7295	7.4357	-
	Absent	4.5415	4.3162	5.0881
Steelhead	Depressed	1.4503	-	
	Migrant	6.9372	6.7879	-
	Absent	3.1854	3.2813	5.6573

Figure captions

Fig. 1. An example of the recursive partitioning process used for tree classification. The trees (top) correspond to their respective graphs (below). The initial partition (left) is at X=30 with the corresponding tree decision if $X \leq 30$ go left. The second partition is at Y = 20 with the corresponding tree decision if $Y \leq 20$ go left. Partitions are separated by broken lines and are labeled with their corresponding tree node identifiers (t). Non-terminal nodes are represented by ovals and terminal nodes by boxes.

Fig. 2. A simplified example of the classification of unknown observations, U1 and U2, as members of one of two groups, A or B. Arrows represent the distance from the unknown observations to their nearest neighbors. Using a $K = 1$ nearest neighbor classification rule (solid arrows), unknown observations U1 and U2 would be classified as members of groups A and B, respectively. A $K = 6$ nearest neighbor rule (all arrows), however, would classify U1 and U2 as members of groups B and A, respectively.

Fig. 3. A graphical representation of a modular neural network with 2 covariate variables, 2 responses, and 2 hidden nodes per module labeled as L_{jk} with j = module and k = hidden node number, respectively. Nodes with B subscripts represent the bias term for the output layer, which is analogous to an intercept in generalized linear models.

Fig 4. (a) Mean p-values from Monte Carlo hypothesis tests of k-nearest neighbor models as a function of jackknife sample size with 500 jackknife samples and (b) the coefficient of variation for p-values as a function of the number of jackknife samples with jackknife sample sizes approximately 15% of the totals. Means and coefficients of variation are from 100 replicate tests of *drnden*, *hk*, and *bank* for Yellowstone cutthroat trout (heavy solid line), redband trout (broken line), and steelhead (thin line), respectively. Jackknife sample sizes are expressed as a percentage of total sample size for each salmonid.

Fig. 5. Examples of estimating optimal values (arrows) of (a) the number of classification tree nodes, (b) the number of K-nearest neighbors, and (c) the number of hidden nodes for the modular neural networks using the overall $\hat{E}ER$.

Fig. 6. A graphical representation of the classification tree for ocean-type chinook salmon population status. Non-terminal nodes are labeled with covariate and number of observations (n) and terminal nodes with predicted status and the distribution of responses in the order: strong, depressed, migrant, and absent. Split-values are to the right of the covariates with node decision: if yes, then down.

Fig 7. A graphical representation of the classification tree for Yellowstone cutthroat trout population status. Non-terminal nodes are labeled with covariate and number of observations (n) and terminal nodes with predicted status and the distribution of responses in the order: strong, depressed, and absent. Split-values are to the right of the covariates with node decision: if yes, then down.

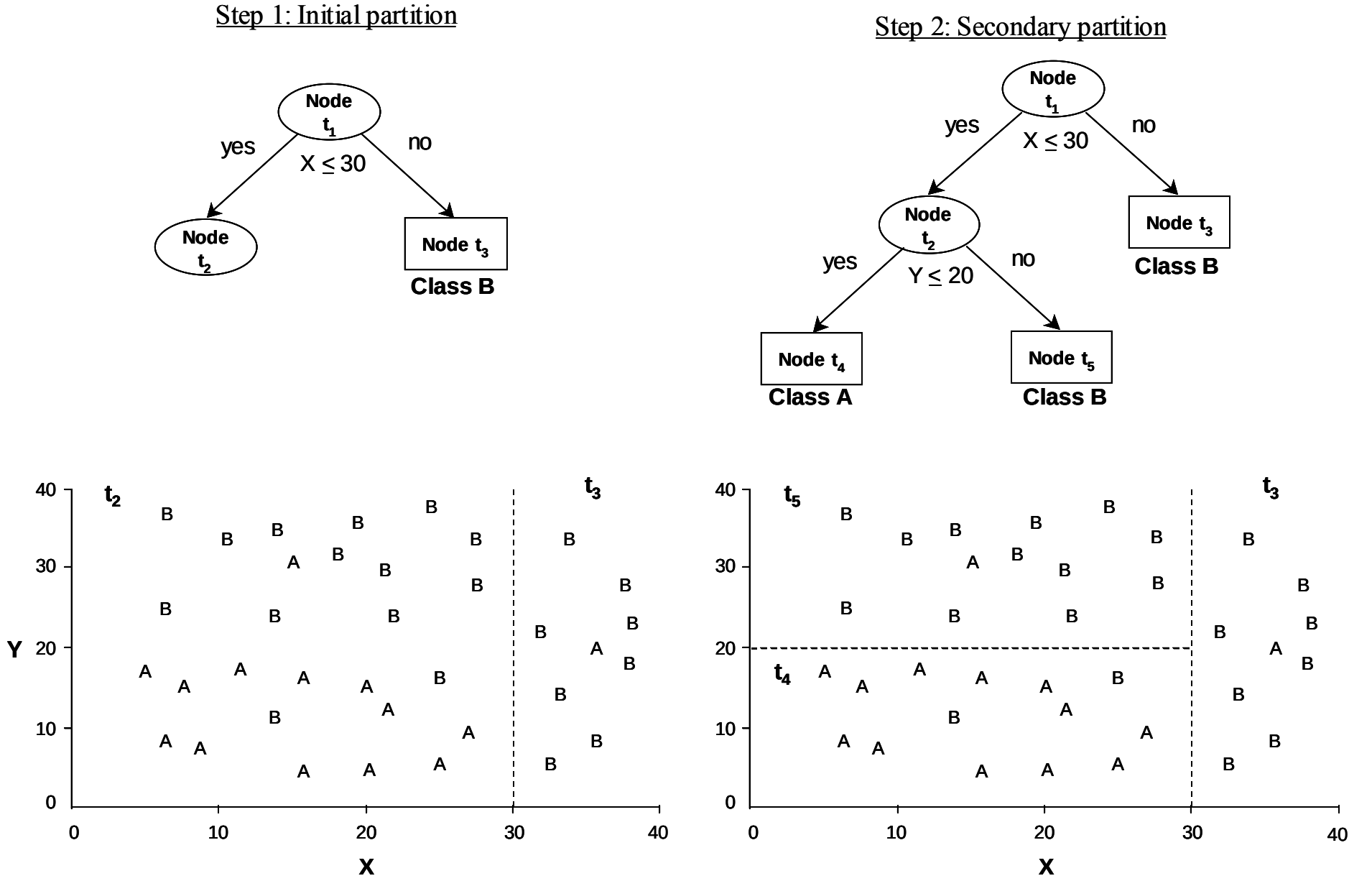
Step 1: Initial partition
Node t1
yes
X ≤ 30
no
Node t2
Node t3
Class B
Step 2: Secondary partition
Node t1
yes
X ≤ 30
no
Node t2
Node t3
Class B
yes
Y ≤ 20
no
Node t4
Class A
Node t5
Class B
t2
t3
t5
t4
Y
X
0
10
20
30
40
A
B

Fig. 1

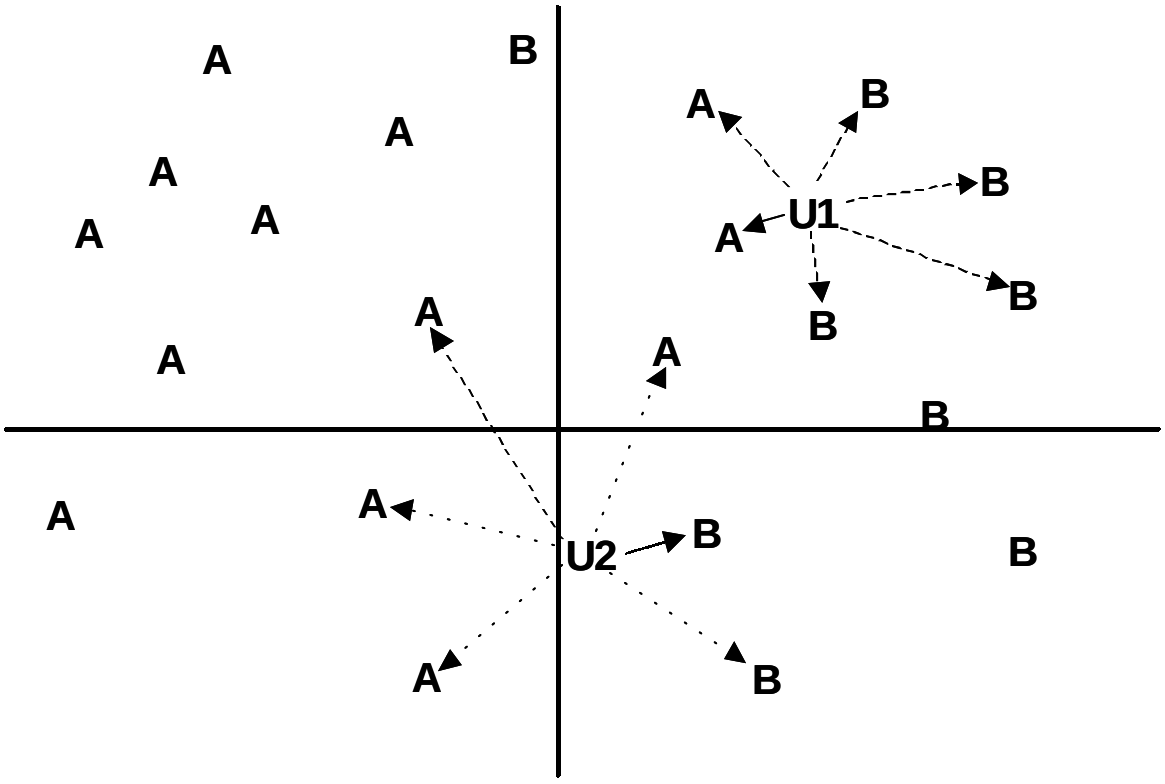

Fig. 2

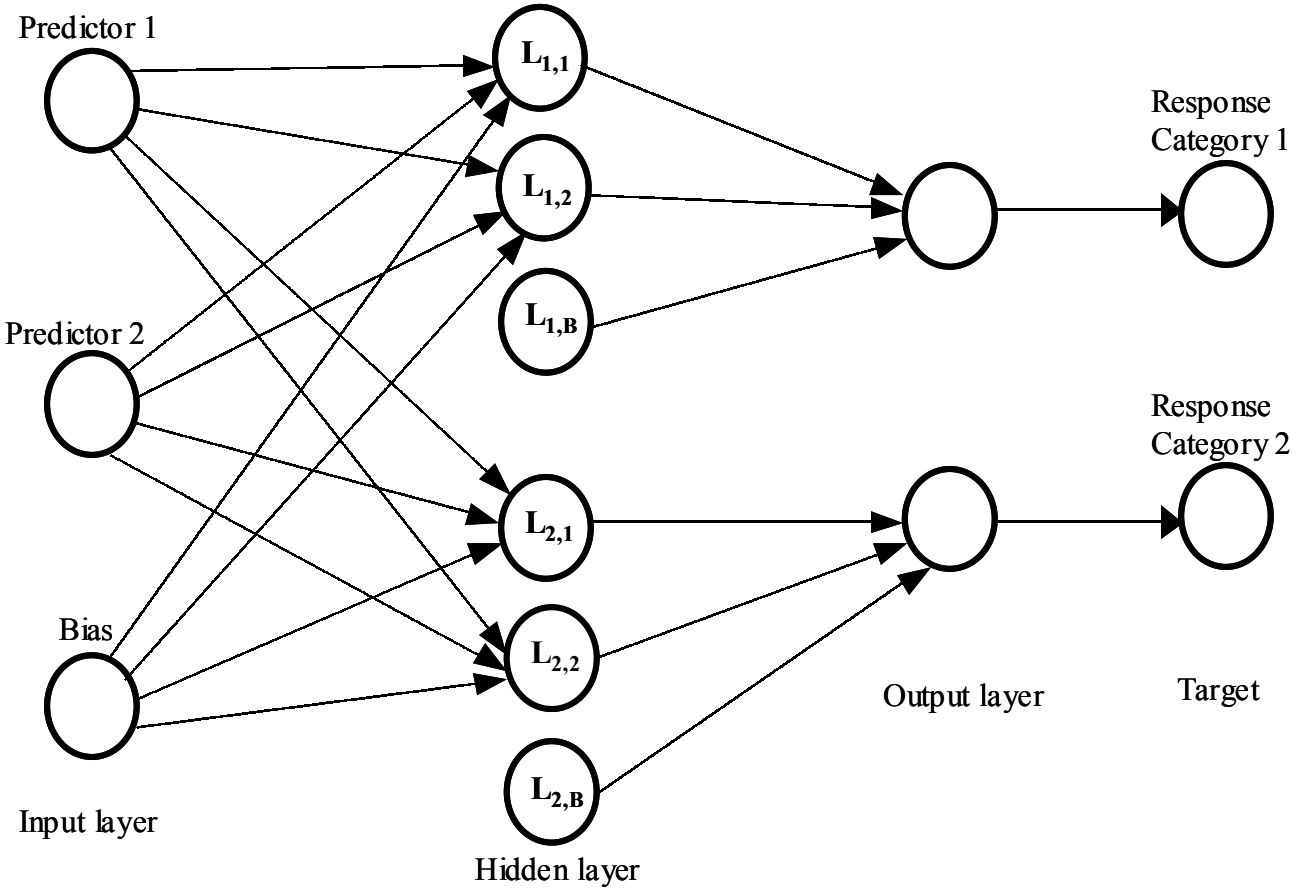
Predictor 1
Predictor 2
Bias
Input layer
$L_{1,1}$
$L_{1,2}$
$L_{1,B}$
$L_{2,1}$
$L_{2,2}$
$L_{2,B}$
Hidden layer
Output layer
Response Category 1
Response Category 2
Target

Fig. 3

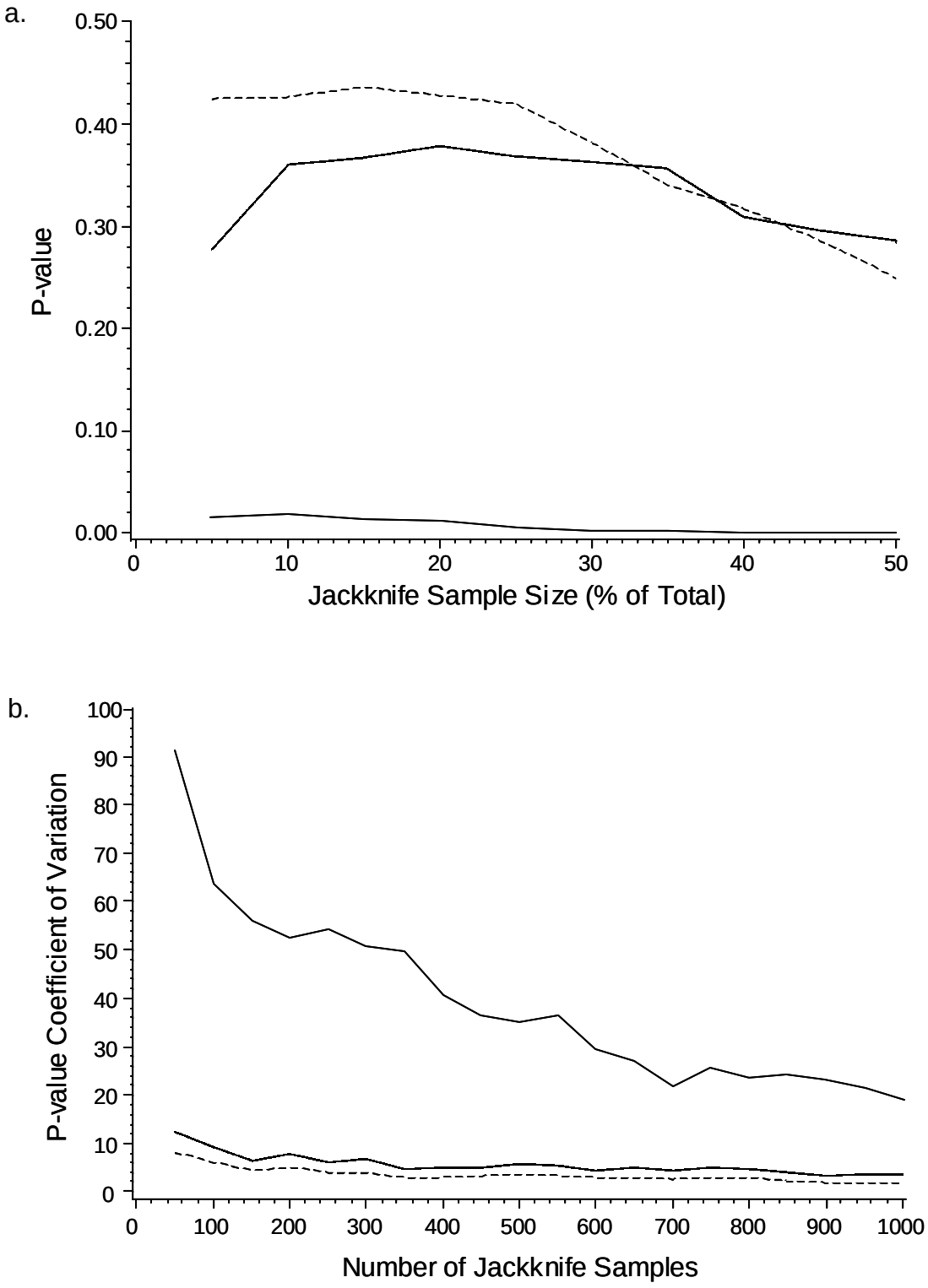
a.
P-value
0.50
0.40
0.30
0.20
0.10
0.00
0
10
20
30
40
50
Jackknife Sample Size (% of Total)
b.
P-value Coefficient of Variation
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
700
800
900
1000
Number of Jackknife Samples

Fig. 4

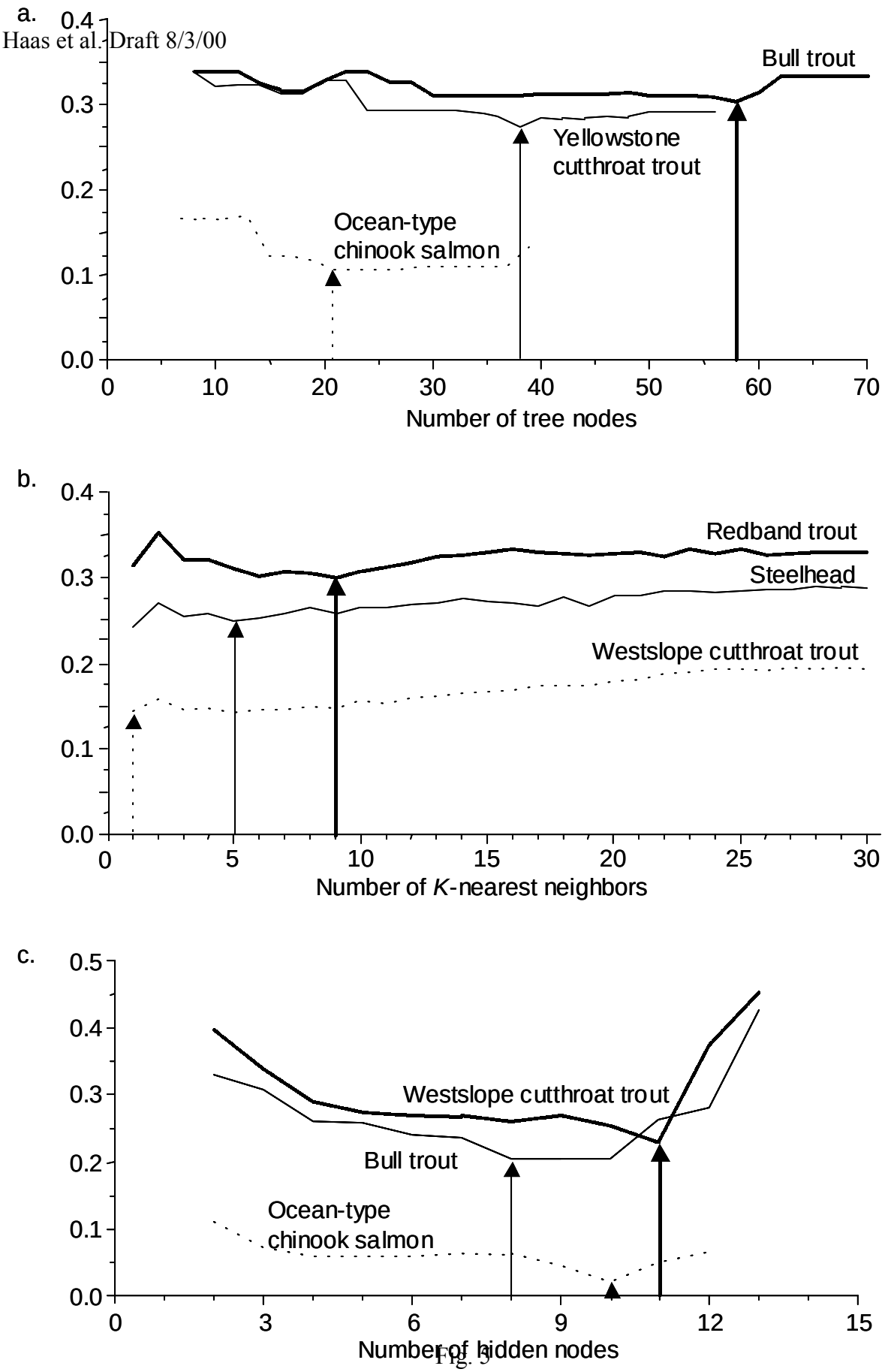
a.
Bull trout
Yellowstone cutthroat trout
Ocean-type chinook salmon
Number of tree nodes
b.
Redband trout
Steelhead
Westslope cutthroat trout
Number of K-nearest neighbors
c.
Westslope cutthroat trout
Bull trout
Ocean-type chinook salmon
Number of hidden nodes

Fig. 5

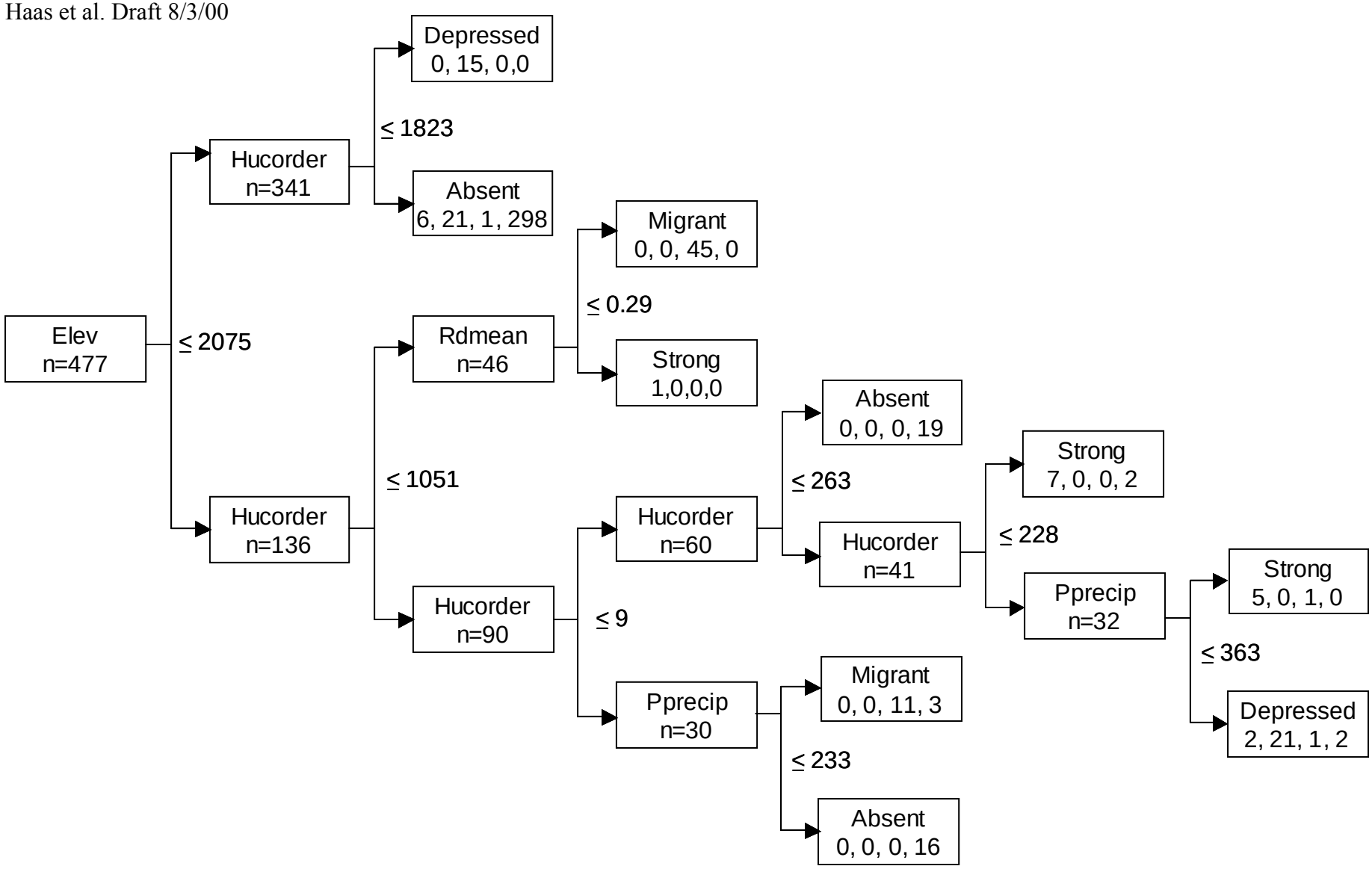
Elev
n=477
≤ 2075
Hucorder
n=341
≤ 1823
Depressed
0, 15, 0,0
Absent
6, 21, 1, 298
Hucorder
n=136
≤ 1051
Rdmean
n=46
≤ 0.29
Migrant
0, 0, 45, 0
Strong
1,0,0,0
Hucorder
n=90
≤ 9
Hucorder
n=60
≤ 263
Absent
0, 0, 0, 19
Pprecip
n=30
≤ 233
Migrant
0, 0, 11, 3
Absent
0, 0, 0, 16
Hucorder
n=41
≤ 228
Strong
7, 0, 0, 2
Pprecip
n=32
≤ 363
Strong
5, 0, 1, 0
Depressed
2, 21, 1, 2

Fig. 6

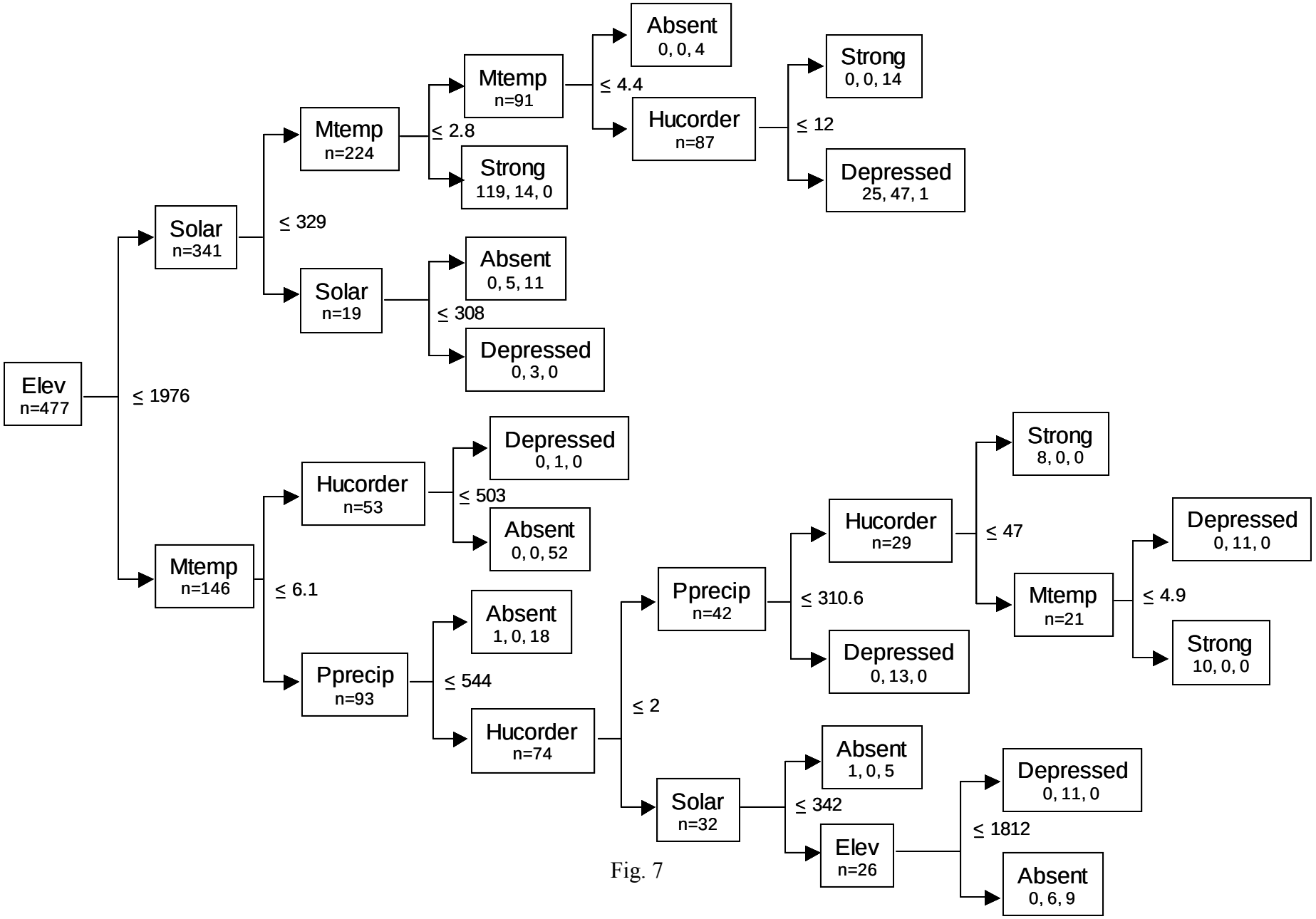
Elev
n=477
≤ 1976
Solar
n=341
≤ 329
Mtemp
n=224
≤ 2.8
Mtemp
n=91
≤ 4.4
Absent
0, 0, 4
Hucorder
n=87
≤ 12
Strong
0, 0, 14
Depressed
25, 47, 1
Strong
119, 14, 0
Solar
n=19
≤ 308
Absent
0, 5, 11
Depressed
0, 3, 0
Mtemp
n=146
≤ 6.1
Hucorder
n=53
≤ 503
Depressed
0, 1, 0
Absent
0, 0, 52
Pprecip
n=93
≤ 544
Absent
1, 0, 18
Hucorder
n=74
≤ 2
Pprecip
n=42
≤ 310.6
Hucorder
n=29
≤ 47
Strong
8, 0, 0
Depressed
0, 13, 0
Mtemp
n=21
≤ 4.9
Depressed
0, 11, 0
Strong
10, 0, 0
Solar
n=32
≤ 342
Absent
1, 0, 5
Elev
n=26
≤ 1812
Depressed
0, 11, 0
Absent
0, 6, 9

Fig. 7