Transactions of the American Fisheries Society 130:878–897, 2001

Fish–Habitat Relationships in Lakes: Gaining Predictive and Explanatory Insight by Using Artificial Neural Networks

JULIAN D. OLDEN*[1] AND DONALD A. JACKSON

Department of Zoology, University of Toronto, Toronto, Ontario, M5S 3G5, Canada

Abstract.—Understanding and predicting the impacts of habitat modification and loss on fish populations are among the main challenges confronting fisheries biologists in the new millennium. Statistical models play an important role in this regard, providing a means to quantify how environmental conditions shape contemporary patterns in fish populations and communities and formulating this knowledge in a framework where future patterns can be predicted. Developing fish–habitat models by traditional statistical approaches is problematic because species often exhibit complex, nonlinear responses to environmental conditions and biotic interactions. We demonstrate the value of a robust statistical technique, artificial neural networks, relative to more traditional regression techniques for modeling such complexities in fish–habitat relationships. Using artificial neural networks, we provide both explanatory and predictive insight into the whole-lake and within-lake habitat factors shaping species occurrence and abundance in lakes from southcentral Ontario, Canada. The results show that species presence or absence is highly predictable based on whole-lake measures of habitat, and that these fish–habitat models show good generality in predicting occurrence in other lakes from an adjacent drainage. Detailed evaluation of these models shows that partitioning the predictive performance of the models into measures such as sensitivity (ability to predict species presence) and specificity (ability to predict species absence) allows assessment of the strengths, weaknesses, and applicability of the models more readily. We show that artificial neural networks are a useful approach for examining the interactive effects of habitat and biotic factors that shape species occurrence, abundance, and spatial occupancy within lakes. Finally, using simulated and empirical examples, we show that artificial neural networks provide greater predictive power than do traditional regression approaches for modeling species occurrence and abundance.

In recent years several broad-scale studies have identified modification and loss of aquatic habitat as primary factors threatening the conservation of freshwater fish populations and communities (Williams et al. 1989; Allen and Flecker 1993; Richter et al. 1997). Consequently, efforts to understand the linkage between habitat, its use by fish, and associated productivity have become increasingly important and currently are central issues in the aquatic sciences (Hughes and Noss 1992; Harig and Bain 1998). Anthropogenic activity has altered many components of riparian areas and nearshore habitats (Jennings et al. 1999). Modifications include changes in the composition and density of macrophytes (Bryan and Scarnecchia 1992), quantity and diversity of shoreline habitat such as woody material (Christensen et al. 1996), and substrate composition (Beauchamp et al. 1994; Jennings et al. 1996). Alterations to littoral-zone habitat can have dramatic and persistent impacts on fish assemblages because this habitat ultimately provides the template on which lentic ecosystems are organized (Jackson and Harvey 1989; Tonn et al. 1990; Hinch et al. 1991; Jackson et al. 2001; Olden et al. 2001).

The ability to evaluate the effects of habitat change and other human impacts on fish populations requires extensive surveying of the fish populations before and after the change occurs (Lester et al. 1996). However, pollution, shoreline development, and other forms of habitat degradation often are not single events for which timing and magnitude are controllable. Such events commonly impose cumulative impacts on fish populations. Indeed, individual effects on populations may be so small relative to natural population variability that statistically significant effects might be detectable only after many years of study. Predictive fish–habitat models may play a useful role in this regard by providing the ability to forecast both small- and large-scale effects of habitat modification on fish populations and communities. For instance, fish–habitat models could provide resource managers with the ability to predict species

* Corresponding author: olden@lamar.colostate.edu

[1] Current address: Graduate Degree Program in Ecology, Department of Biology, Colorado State University, Fort Collins, Colorado 80523–1878, USA.

Received June 2, 2000; accepted March 23, 2001

occurrence and abundance at different spatial scales by using whole-lake and within-lake measures of habitat. Ultimately, predictive models would enhance managers' abilities to predict the temporal and spatial scales at which habitat can be changed while minimizing the impact to lake fish populations.

Although fish–habitat models play an important role in fisheries ecology and management, developing useful models may be difficult because species exhibit complex, nonlinear responses to environmental and biotic factors. Multiple linear regression and linear discriminant analyses remain the most frequently used techniques for modeling fish–habitat relationships, although our confidence in these methods is often limited by the inability to meet some of the assumptions of the model, such as appropriate error structure of the variables, independence of variables, and model linearity (James and McCulloch 1990). The last assumption is particularly susceptible to violation when ecological data are examined. Data transformations of variables can improve the results of traditional approaches, but often these are only partially successful (e.g., Lek et al. 1996; Guégan et al. 1998; Wally and Fontama 1998). Furthermore, the choice of transformation may influence the results and thus bias our interpretation of ecological relationships.

Artificial neural networks (ANNs) are a promising alternative to traditional statistical approaches, providing a powerful, flexible learning technique for uncovering nonlinear patterns in data. Applications of ANNs are diverse in the literature, ranging from social sciences to chemistry, and recently have received more attention in the ecological sciences for their ability to solve complex pattern-recognition problems (Colasanti 1991; Edwards and Morse 1995; Lek et al. 1996). ANNs can have advantages over traditional methods when applied to systems that may be poorly defined and understood and to situations where input data are incomplete or ambiguous by nature. Furthermore, unlike the more commonly used methods, neural networks are not dependent on particular functional relationships, make no assumptions regarding the distributional properties of the data, and require no a priori understanding of variable relationships. This independence makes ANNs a potentially powerful modeling tool for exploring complex, nonlinear biological problems, such as the relationships believed to exist between fish and their surrounding environment.

The primary objectives of our study are to highlight the value of ANNs for modeling ecological relationships and to illustrate their ability to provide insight into understanding and predicting relationships between fish populations and the environment. Before addressing these objectives, however, we provide a simple methodological comparison between ANNs and traditional regression-based approaches in predicting simulated patterns of species presence/absence and abundance relative to an environmental gradient. The results from this comparison demonstrate the capabilities of these various approaches under deterministic or known conditions. We then use field data to model the relationships between lakewide habitat attributes and species occurrence in a set of temperate lakes located in southcentral Ontario, Canada. More specifically, we determine the predictability of species presence/absence on the basis of readily available, whole-lake habitat factors (e.g., surface area, maximum depth, elevation) and go beyond conventional model evaluations by estimating optimal decision thresholds for prediction to maximize measures of classification success, sensitivity, and specificity of the models. Next, we test the performance of these models for predicting species occurrences from a second set of lakes in an adjacent drainage, providing an assessment of the transferability or generality of the species–habitat models. Given that species abundance may be a more sensitive response variable for studying fish–habitat relationships (but see Jackson and Harvey 1997), we model associations between within-lake species abundances and nearshore habitat features (e.g., macrophyte cover, substrate types, site exposure) for several littoral-zone fishes. Finally, we empirically compare the predictive performance of the neural networks to that of fish–habitat models developed by using logistic regression for species occurrence and multiple regression for species abundance. We show that ANNs rovide powerful predictive models and shed important insight into the individual and interactive relationships between fish and their environment at local and regional scales.

Artificial Neural Networks

The ability of the human brain to perform complex tasks, such as pattern recognition, has motivated a large body of research exploring the computational capabilities of highly connected networks of relatively simple elements, ANNs. Although ANNs were initially developed to better understand how the mammalian brain functions, researchers in a variety of scientific disciplines

have become more interested in the potential mathematical utility of neural network algorithms for addressing an array of problems. For example, ANNs have shown great promise for solving complex pattern-recognition problems and for developing prediction or classification rules in the biological sciences (e.g., Colasanti 1991; Edwards and Morse 1995; Lek et al. 1996; Lek and Guégan 1999). Previous studies using ANNs are too numerous to list here; however, their use in fisheries applications has been limited and includes modeling fish species richness (Guégan et al. 1998), presence/absence (Mastrorillo et al. 1997), abundance (Lek et al. 1996; Brosse et al. 1999), and production (Chen and Ware 1999).

Although many types of ANNs exist (see Bishop 1995; Ripley 1996), here we describe the type used most frequently; the one-hidden-layer, feed-forward neural network trained by the back-propagation algorithm (Rumelhart et al. 1986). These extremely popular neural networks have been used in the biological literature because they are considered to be universal approximators of any continuous function (Cybenko 1989; Funahashi 1989; Hornick et al. 1989). Furthermore, single hidden-layer networks greatly reduce computational time and often produce results similar to those obtained by multiple hidden-layer networks (Kurková 1992). Below, we discuss two important features of ANNs: network architecture and the back-propagation algorithm used to parameterize the network, and interpretation of variable importance in the network.

Network architecture and the back-propagation algorithm.—Network architecture refers to the number and organization of the computing units (called neurons) in the network. In the one-hidden-layer, feed-forward network, neurons are organized in an input layer, a hidden layer, and an output layer, with each layer containing one or more neurons (Figure 1). Each neuron is connected by an axon to all neurons in adjacent layers; however, neurons within each layer and in nonadjacent layers are not connected. The input layer typically contains p neurons, one neuron representing each of the predictor variables $x_1 \ldots x_p$. The number of neurons in the hidden layer is determined empirically by the investigator to minimize the trade-off between bias and variance (Geman et al. 1992). Additional hidden neurons increase the ability of a network to approximate any underlying relationship among the variables, that is, reduced bias, but result in the network having a large number of free parameters, thereby increasing the variance

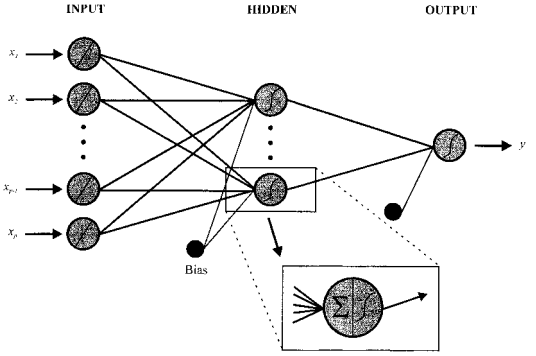

FIGURE 1.—A one-hidden-layer, feed-forward neural network design.

of predictions because of overfitting the data. Although mathematical derivations exist for selecting an optimal design, in practice it is common to train networks with different numbers of hidden neurons and use the performance on a test set to choose the network that performs best. For continuous and binary response variables the output layer commonly contains one neuron. However, the number of output neurons can be greater than one if there is more than one response variable or if the response variable is categorical (that is, a separate neuron is used for classifying observations into each category). Additional neurons with a constant output (commonly set to 1) are also added to the hidden and output layers (Figure 1), although this inclusion is not mandatory. These are called bias neurons, and play a role similar to that of the constant term in multiple regression analysis.

The connection between any two neurons is assigned a weight that dictates the intensity of the signal they transmit through the axon. Consequently, the ''state'' or ''activity level'' of each neuron is determined by the input received from the other neurons connected to it. In feedforward networks, axon signals are transmitted in a unidirectional path from input layer to output layer through the hidden layer. The states of the input neurons are defined by the incoming signal or values of the predictor variables. The state of each hidden neuron is evaluated locally by calculating the weighted sum of the incoming signals from the neurons of the input layer (Figure 1 inset) and then adding a bias input. The weighted sum is then subjected to an activation function, that is, a differentiable function of the neuron's total incoming signal from the input neurons, to produce the state of the hidden neuron (Figure 1 inset). The same

procedure described above is repeated for the axon signals from the hidden layer to the output layer. The entire process can be written mathematically as

$$y_k = \phi_o\left\{\beta_k + \sum_j w_{jk}\phi_h\left(\beta_j + \sum_i w_{ij}x_i\right)\right\} \quad (1)$$

where x_i are the input signals, y_k are the output signals, w_{ij} are the weights between input neuron i to hidden neuron j, w_{jk} are the weights between hidden neuron j and output neuron k, β_j and β_k are the biases associated with the hidden and output layers, respectively, and ϕ_h and ϕ_o are activation functions for the hidden and output layers, respectively. There are several activation functions, but the logistic function defined as

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

is the most commonly used.

Training the neural network involves an error back-propagation algorithm, which finds a set of connection weights that produces an output signal with a small error relative to the observed output. During training, the weights are adapted to minimize some fitting criterion. For continuous output variables, the most commonly used criterion is the least-squares error function

$$E = \sum_n \|t^n - y^n\|^2 \quad (3)$$

whereas for dichotomous output variables, the most commonly used criterion is the cross-entropy (i.e., similar to log-likelihood) error function (Bishop 1995):

$$E = -\sum_n \{t^n \log_e y^n + (1 - t^n)\log_e(1 - y^n)\} \quad (4)$$

where t^n is the observed output value and y^n is the predicted output value for the observation. The algorithm adjusts the connection weights in a backwards fashion, layer by layer, in the direction of steepest descent in minimizing the error function (also called gradient descent). One iteration of the gradient descent algorithm can be summarized as follows:

$$\Delta w_{st} = -\frac{\partial E}{\partial w_{st}} \quad (5)$$

where Δw_{st} is the weight change between neuron s and neuron t in the next layer. The training of the network is a recursive process in which observations from the training data are entered into the network in turn, each time modifying the input–hidden and hidden–output connection weights (using equation 5). This procedure is repeated with the entire training data set (i.e., each of the n observations) for several iterations until a stopping rule is achieved. This type of training is a sequential approach to network optimization and is in contrast to the batch approach, in which the entire data set is used to adjust the weights during each iteration (Bishop 1995). Commonly, network training is stopped when the difference between predicted outputs from the network and the observed output (i.e., the error function) is small or when the possibility of overfitting the data is minimized.

Interpreting variable importance in ANNs.—Although many studies have shown ANNs exhibit greater predictive power than traditional approaches do (e.g., Lek et al. 1996), researchers often call it a ''black box'' approach to statistical modeling because the networks are believed to provide little explanatory insight into the relative influence of the independent variables in the prediction process (Lek and Guégan, 1999; Özesmi and Özesmi, 1999). The lack of explanatory power is a major concern because the interpretation of statistical models is desirable for gaining knowledge of the causal factors driving ecological phenomena. This has been a major pitfall of ANNs because traditional statistical approaches can readily identify the influence of the independent variables in the modeling process and also provide a degree of confidence regarding their contribution. Fortunately, recent studies have provided various methods for quantifying and interpreting the contributions of the independent variables in neural networks. For example, several intensive computational approaches have been developed, including growing and pruning algorithms (Bishop 1995), partial derivatives (e.g., Dimopoulos et al. 1995), and asymptotic t-tests.

In the neural network, the connection weights between neurons are the linkages between the inputs and the output of the network and therefore are the link between the problem and the solution. Consequently, the relative contribution of each independent variable to the predictive output of the neural network depends primarily on the magnitude and direction of these connection weights. Input variables with larger connection weights represent greater intensities of signal transfer and therefore are more important in predicting the output than are variables with smaller weights. Neg-

ative connection weights represent inhibitory effects on neurons (reducing the intensity or contribution of the incoming signal and negatively affecting the output), whereas positive connection weights represent excitatory effects on neurons (increasing the intensity of the incoming signal and positively affecting the output). Recently, some studies have used connection weights to interpret the participation of input variables in predicting the output of the network (e.g., Aoki and Komatsu 1997; Chen and Ware 1999; Özesmi and Özesmi 1999). Other approaches involve using all the weights of the network to quantify overall variable importance (e.g., Garson 1991) and use sensitivity analysis to determine the spectrum of input variable contributions in the neural network (e.g., Lek et al. 1996; Mastrorillo et al. 1997; Guégan et al. 1998). Although these approaches can determine the overall influence of each predictor variable, the interpretation of interactions among the variables is more difficult to assess because the strength and direction of individual axon connection weights within a network must be examined directly. Even with small networks, the number of connections is large, and thus the interpretation of the network is difficult. For example, a network containing 10 input neurons and 7 hidden neurons would have 70 connection weights to examine. Bishop (1995) suggested removing small weights from the network to ease interpretation; however, deciding which weights should be retained or eliminated from the network is a difficult task.

We have developed a randomization test for ANNs to address this question. This approach randomizes the response variable, then constructs a neural network by using the randomized data and records all input—hidden–output connection weights (the product of the input–hidden and hidden–output weights). This process is repeated 10,000 times to generate a null distribution for each input–hidden–output connection weight (10,000 randomizations ensures stability of the estimated probability values; Jackson and Somers 1989), which is then compared with the observed values to calculate the significance level (see Olden 2000a for more details). The randomization test provides an objective pruning technique for eliminating connection weights that have minimal influence on the network output and identifies independent variables that significantly contribute to the prediction process.

In this study, the optimal number of neurons in the hidden layer was determined empirically by comparing the performance of different networks having 1–20 hidden neurons and choosing the network with the best predictive performance. We included learning rate (η) and momentum (α) parameters (which vary as a function of the network prediction error) during network training to ensure a high probability of global network convergence (Bishop 1995) and considered a maximum of 1,000 iterations for the back-propagation algorithm to determine the optimal axon weights. Before training the neural network, the independent variables for both modeling species occurrence and abundance were converted to *z*-scores to standardize the measurement scales of the inputs into the network and thus ensure that the same percentage change in the weighted sum of the inputs caused a similar percentage change in the unit output.

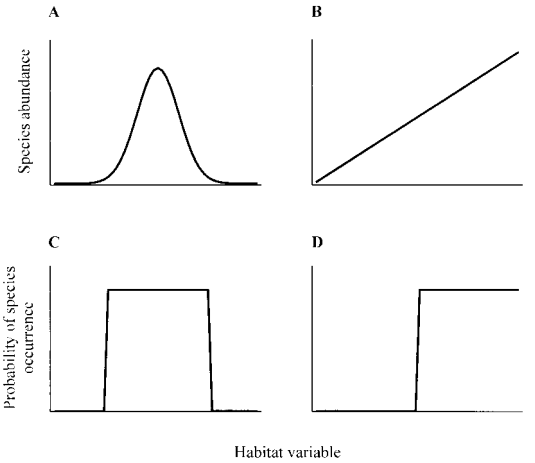

FIGURE 2.—Functional response of simulated data sets showing **(A, B)** species abundance and **(C, D)** presence/absence response curves to a single habitat variable. See Methods for description of data.

Methods

ANNs versus traditional regression approaches: simulated data.—We compared the predictive performance of ANNs and regression models by using simulated data with defined correlations among the variables. To compare the performance of neural networks and linear regression models for predicting species abundance (i.e., continuous response variable), we generated two data sets, each containing 50 observations and with values of the habitat variable ranging from 0 to 10. The first data set has a Gaussian or normal-curve response (characterized by a mean of 5 and a variance of 1) in the abundance of a species to the habitat variable (Figure 2a). This response curve is commonly used in theoretical models that are related

to the species niche concept. In the second data set, species abundance was generated to show a linear association (slope = 1) to the habitat variable (Figure 2b). Similarly, two data sets were generated to compare ANNs and logistic regression analysis for predicting species presence/absence (i.e., dichotomous response variable). The first data set was constructed by converting the Gaussian data set described above (Figure 2a) to a presence/absence response distribution (Figure 2c). The conversion was accomplished by defining all abundances greater than 0.02 to represent species presence; those below this value were coded as species absence. As a result, this data set contained an equal number of presence and absence values. In the second data set, the probability of species occurrence was simulated to have a logistic (or sigmoidal) response curve, following equation (2), with the habitat variable ranging from −5 to 5. Using the simulated data sets described above, we compared the predictive performance of the neural networks with that of the traditional regression models. The linear species abundance curve and logistic species occurrence curve represent ''optimal'' data types for the traditional regression approaches in terms of distributional characteristics, whereas the Gaussian species abundance and occurrence curves represent a nonlinear relationship in which the probability of species occurrence or abundance is maximized at intermediate values of a habitat variable (e.g., the probability of a species occurrence in relation to stream velocity such that the species is absent from pools, abundant within moderately flowing water, but absent from high-velocity systems). For all analyses the data points were sampled at uniform distances along the habitat variable.

ANNs for fish species presence/absence.—The study sites were 128 lakes from the Madawaska River drainage and 32 lakes from the Oxtongue River drainage, located in Algonquin Provincial Park, Ontario, Canada (45°50'N, 78°20'W; Figure 3a). Aquatic communities in this region are representative of relatively natural ecosystems because the lakes are located in a provincial park and are currently subject to minimal perturbations from human development—although limited species introductions (e.g., smallmouth bass) were made into the area during the early 1900s, which subsequently colonized some adjacent waters. We developed fish–habitat models for nine fish species—brown bullhead *Ameiurus nebulosus*, common shiner *Luxilus cornutus*, creek chub *Semotilus atromaculatus*, golden shiner *Notemigonus crysoleucas*, lake trout *Salvelinus namaycush*, northern redbelly dace *Phoxinus eos*, pumpkinseed *Lepomis gibbosus*, smallmouth bass *Micropterus dolomieu*, and yellow perch *Perca flavescens*—by modeling species presence/absence as a function of seven whole-lake variables. The predictor variables are factors related to known habitat requirements of fish in this geographic region (Matuszek and Beggs 1988; Minns 1989): surface area, total shoreline perimeter, maximum depth, total dissolved solids (TDS), pH, lake elevation, and occurrence of summer stratification (Table 1). For small-bodied fish (i.e., common shiner, creek chub, golden shiner, and northern redbelly dace), we included the presence/absence of a littoral-zone predator (smallmouth bass, largemouth bass, or northern pike) as an additional predictor variable because predation could be an important force (Jackson et al. 2001). Data were obtained from the Algonquin Park Fish Inventory Data Base (Crossman and Mandrak 1991); details of sampling methodologies are as described in Dodge et al. (1985).

To evaluate predictive performance, we validated the fish–habitat models by using two approaches. First, we used *n*-fold or ''leave-one-out'' cross-validation (also referred to as jackknife validation) to assess model performance in 128 lakes from the Madawaska River drainage. This technique provides a nearly unbiased estimate of model performance (Olden and Jackson 2000). Second, we tested the ability of the Madawaska-drainage models to predict species occurrence in 32 lakes from the adjacent Oxtongue River drainage. This analysis provides an opportunity to assess the transferability or generalization of the models to other drainages in the same geographic region. We partitioned the overall classification success of each species model by deriving confusion matrices (Fielding and Bell 1997). A confusion matrix tabulates the observed and predicted presence/absence patterns and thus provides a summary of the number and direction of correct and incorrect classifications produced by the model. Using these matrices, we examined three metrics of prediction success. First, we quantified the overall classification performance of the model as the percentage of lakes where the model correctly predicts the presence/absence of the species (correct classification). Second, we examined the ability of the model to accurately predict species presence (model sensitivity). Third, we examined the ability of the model to accurately predict species absence (model specificity). Rather than simply following the conventional decision threshold of 0.5 (the cut-

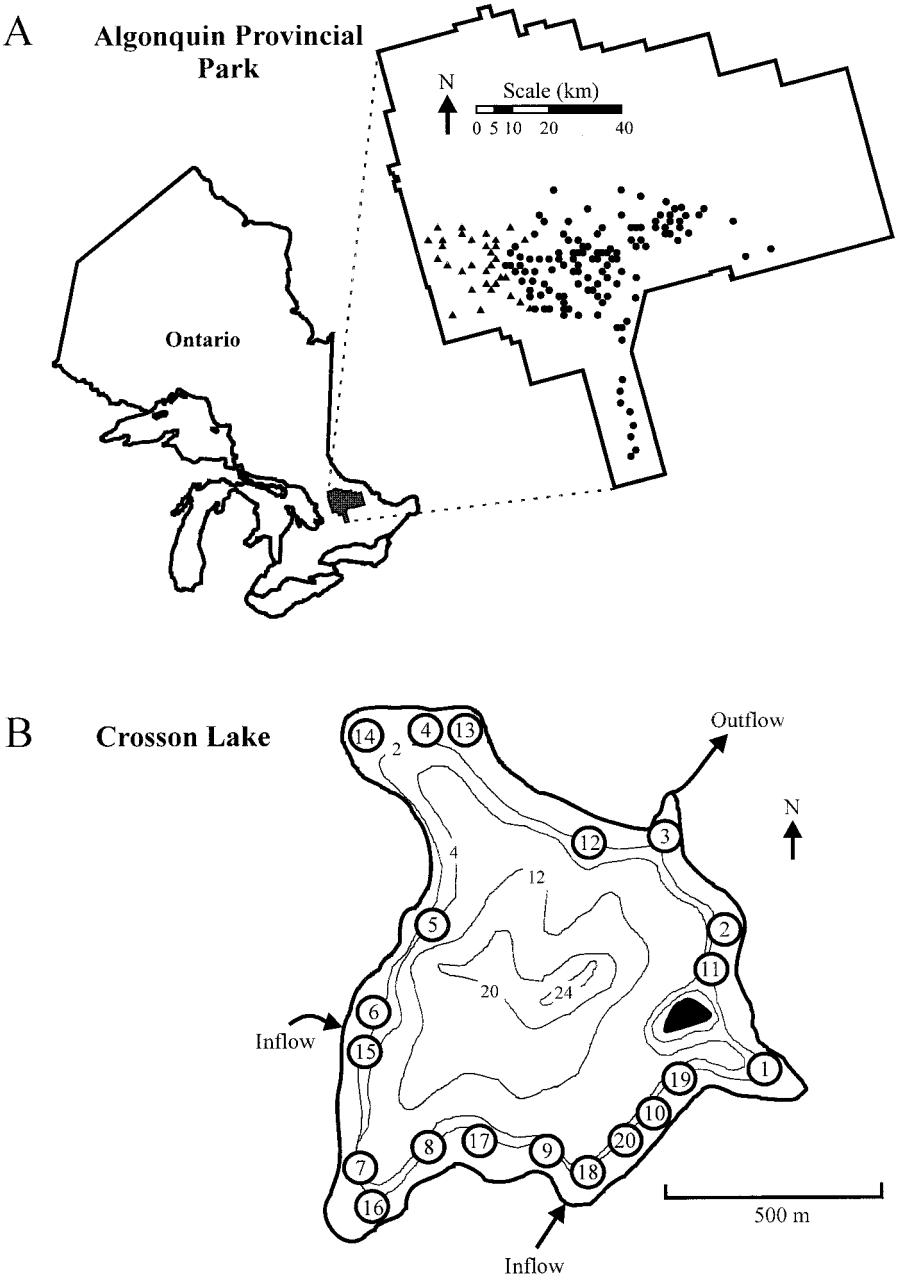

FIGURE 3.—**(A)** Location of study lakes from the Madawaska River drainage (128 lakes depicted by circles) and Oxtongue River drainage (32 lakes depicted by triangles) in Algonquin Provincial Park, Ontario, Canada (45°50'N, 78°20'W), and **(B)** Crosson Lake (45°05'N, 77°20'W), with 20 sampling stations depicted by numbered circles.

off at which a species is predicted to be present), we constructed receiver-operating characteristic (ROC) plots for each species to estimate the predictive ability of the models over all decision thresholds (Metz 1978; Fielding and Bell 1997). A ROC graph is a plot of the sensitivity/specificity pairs resulting from continuously varying the decision threshold over the entire range of results observed. The optimal decision threshold is chosen to maximize overall classification performance of the model, assuming the costs of misclassifying the species as present or absent are equal. The

TABLE 1.—Summary statistics of whole-lake habitat variables used in the neural networks and logistic regression models to predict species presence/absence. The abbreviation CI stands for confidence interval.

	Madawaska River drainage (training data)		Oxtongue River drainage (test data)	
Whole-lake variables	Median	95% CI	Median	95% CI
Area (ha)	29.6	(4.0, 417.6)	83.5	(10.1, 617.1)
Maximum depth (m)	13.7	(3.1, 36.1)	19.1	(3.9, 49.3)
Shoreline perimeter (km)	4.0	(1.0, 22.9)	8.0	(2.0, 43.1)
Elevation (m)	432	(390, 475)	436	(419, 488)
Total dissolved solids (mg/L)	26.0	(18.4, 51.3)	22.0	(18.1, 48.3)
pH	7.0	(6.0, 7.5)	7.0	(6.3, 7.5)
Summer stratification (0, 1)				
Littoral-zone predator (0, 1)				

optimal decision threshold is used to calculate correct classification, sensitivity, and specificity, and Cohen's kappa statistic is used to assess whether the performance of the model differs from expectations based on chance alone (Titus et al. 1984).

Next, we provided an empirical comparison of ANNs with logistic regression analysis for predicting species presence/absence. The predictive performances of ANNs and logistic regression models for species occurrence were compared for rates of overall correct classification, sensitivity, and specificity. Note that ROC analysis was used to determine optimal decision thresholds for the logistic models also. Screening the data characteristics of the habitat variables before analyses led us to use $\log_e(x)$ transformation of all continuous variables except pH and the littoral-zone predator variable.

ANNs for fish species abundance.—In the within-lake analysis we examined fish–habitat associations for four of the most abundant species (golden shiner, creek chub, pumpkinseed, and yellow perch) in Crosson Lake, located in southcentral Ontario, Canada (45°05'N, 79°02'W). Sampling was done twice (in July and August) and consisted of approximately 24-h sets of baited minnow traps at depths of both 0.5 m and 1.5 m at 20 locations around the perimeter of the lake (Figure 3b). The relative abundances of species were calculated by standardizing each catch to a 24-h sampling period. Eight habitat variables were examined. Sites were categorized on the basis of substrate type (categorized into eight ordered categories based on particle size, ranging through muck, clay, silt, sand, gravel, rubble, and boulder to bedrock), substrate diversity (a measure of the diversity of bottom types present), presence of terrestrial leaf litter, relative cover of vegetation (none, sparse, moderate, or dense), relative cover of woody materials (none, sparse, moderate, or dense), degree of exposure (none, limited, moderate, or extreme), depth (0.5 or 1.5 m), and sampling month; sampling month was included as a binary predictor variable to determine whether a temporal component was important in predicting relative abundance. Habitat was assessed visually from within a boat at each sampling location, and sites containing multiple values of any habitat variable were averaged to give a single value per site.

Associations between species abundance and the eight within-lake habitat variables were modeled by using ANNs (again determining the optimal number of hidden neurons between 1 and 20). The dependent variable was standardized to the range from 0 to 1 to conform to the requirements of the logistic transfer function used in building the neural network. Predictive performance of the models was evaluated by using n-fold cross-validation, as was done for the species-occurrence models. Performance of the models was assessed by Pearson's product-moment correlation coefficient between predicted and actual species abundance, and the root-mean-square-of-error (RMSE) of the predicted values. The Pearson's correlation provides a measure of model accuracy, the better models being represented by correlation coefficients approaching 1. RMSE measures model precision, small values representing high precision and large values indicating low precision.

We then provided an empirical comparison of ANNs with linear regression analysis for predicting species abundance. Predictive performances of ANNs and regression models for species abundance were compared for values of Pearson's product-moment correlation coefficients and RMSE. Data characteristics of the habitat variables were screened before analyses, which led us to perform $\log_e(x)$ transformation of all variables and standardization to z-scores.

TABLE 2.—Comparison of neural network and regression-based approaches for modeling simulated species response curves (see Methods for description of data). Reported values are the percent correct classification (CC), sensitivity (SE), and specificity (SP) for modeling species presence/absence and Pearson's correlation coefficient (*r*) between predicted and actual values and the root mean square error (RMSE) of the prediction for modeling species abundance.

	Neural network vs. logistic regression						Neural network vs. linear regression			
	Neural network			Logistic regression			Neural network		Linear regression	
Variable	CC	SE	SP	CC	SE	SP	*r*	RMSE	*r*	RMSE
Presence/absence										
Gaussian species response	90	88	92	0	0	0				
Logistic species response	100	100	100	100	100	100				
Abundance										
Gaussian species response							1.00	0.017	0.10	0.230
Linear species response							1.00	0.000	1.00	0.000

Results

ANNs versus Traditional Regression Approaches: Simulated Data

Examining the simulated species occurrence data (Table 2) makes evident that ANNs greatly outperform logistic models for the Gaussian species response curve (i.e., nonlinear species–habitat relationship). Importantly, in this case the logistic model never resulted in a correct classification because the logistic function was unable to successfully model the species–habitat relationship, and thus predictions were based solely on chance. Given that the simulated data contained equal numbers of species presences and absences, the accuracy of chance predictions from the logistic model will depend directly on the portion of species presence and absence values (i.e., species prevalence) in the data. If during the cross-validation procedure a case of species absence is removed before construction of the model, species presence will be greater than 50% in the data set, and the logistic model therefore will randomly predict species presence even though ultimately this decision is incorrect. This removal thus results in rates of correct classification, sensitivity, and specificity equal to zero. In contrast, where a species occurrence follows a logistic response curve in relation to the habitat variable, both ANNs and logistic models have perfect classification success. This result is expected for the logistic model, given that the specific data assumptions were met. Together, the results show that ANN greatly outperforms the linear regression models for the Gaussian species response curve, whereas both approaches perform equally in the linear case (Table 2).

ANNs for Fish Species Presence/Absence

Whole-lake attributes were found to be useful predictors of species presence/absence (Table 3). Across both drainages, species were classified correctly in 60.9–84.5% of the lakes, whereas model sensitivity and specificity varied widely among species and between drainages. In the Madawaska drainage the predictive performance for seven of the nine species–habitat models differed significantly from random. Smallmouth bass and lake trout exhibited the highest correct classification rates, creek chub and pumpkinseed showed the greatest sensitivity, and brown bullhead and golden shiner had the greatest specificity. The neural interpretation diagrams for smallmouth bass, lake trout, common shiner, and northern redbelly dace are shown in Figure 4. In these diagrams, the relative magnitude of the connection weights is represented by line thickness (i.e., thicker lines representing greater weights) and line type represents the direction of the weights (i.e., solid lines represent positive signals and dashed lines represent negative signals). The relationship between the inputs and outputs is determined in two steps because there are input–hidden layer connections and hidden–output layer connections. Positive effects of input variables are depicted by positive input–hidden and positive hidden–output connection weights, or negative input–hidden and negative hidden–output connection weights. Negative effects of input variables are depicted by positive input–hidden and negative hidden–output connection weights, or by negative input–hidden and positive hidden–output connection weights. Therefore, the multiplication of the two connection weight directions (positive or negative) indicates the effect that each input variable has on the response variable. Interactions among predictor variables can be identified as input variables with opposing connection weights entering the same hidden neuron. The total contribution of an input variable is calculated as the sum of the products of

TABLE 3.—Performance of neural networks in predicting species presence/absence in 128 lakes in the Madawaska River drainage (training data) based on *n*-fold cross validation and application of the Madawaska networks to predict presence/absence in 32 lakes from the Oxtongue River drainage (test data). The reported values are the number of hidden neurons in the network (#HN), percent species occurrence in the drainage (SO), the optimal decision threshold based on receiver-operating characteristic analysis (ODT) in which species with model probabilities greater than this value are predicted to be present, percent correct classification (CC), sensitivity (SE), specificity (SP), and the Kappa *z*-value statistic and associated *P*-value.

		Madawaska River drainage (training data)							Oxtongue River drainage (test data)					
Species	#HN	SO	ODT	CC	SE	SP	Kappa	*P*	SO	CC	SE	SP	Kappa	*P*
Brown bullhead	5	37.5	0.59	67.2	41.7	82.5	2.58	0.005	65.6	65.6	90.5	18.2	0.45	0.326
Common shiner	3	43.8	0.59	66.4	64.3	68.1	3.60	0.000	53.1	84.5	82.4	86.7	3.88	0.000
Creek chub	4	65.6	0.90	65.6	95.2	9.1	0.46	0.323	62.5	78.1	90.0	58.3	2.58	0.005
Golden shiner	3	35.2	0.54	64.8	35.6	80.7	1.70	0.045	37.5	62.5	41.7	75.0	0.89	0.187
Lake trout	2	43.0	0.54	75.0	70.9	78.1	5.18	0.000	56.3	78.1	88.9	64.3	2.96	0.002
Northern redbelly dace	4	53.1	0.53	60.9	58.8	63.3	2.47	0.007	28.1	71.9	55.6	78.3	1.57	0.058
Pumpkinseed	2	60.2	0.79	68.8	94.8	29.4	1.44	0.075	65.6	65.6	90.5	18.2	0.45	0.326
Smallmouth bass	4	25.0	0.39	80.5	50.0	90.6	3.62	0.000	46.9	71.9	53.3	88.2	2.34	0.010
Yellow perch	3	68.0	0.52	72.7	93.1	29.3	2.28	0.011	75.0	81.3	95.8	37.5	1.53	0.063
Mean		47.9		69.1	71.7	54.5	2.59		54.5	73.3	80.4	54.4	1.85	
SD		13.9		6.0	23.3	26.5	1.31		14.1	7.7	18.7	25.1	1.11	

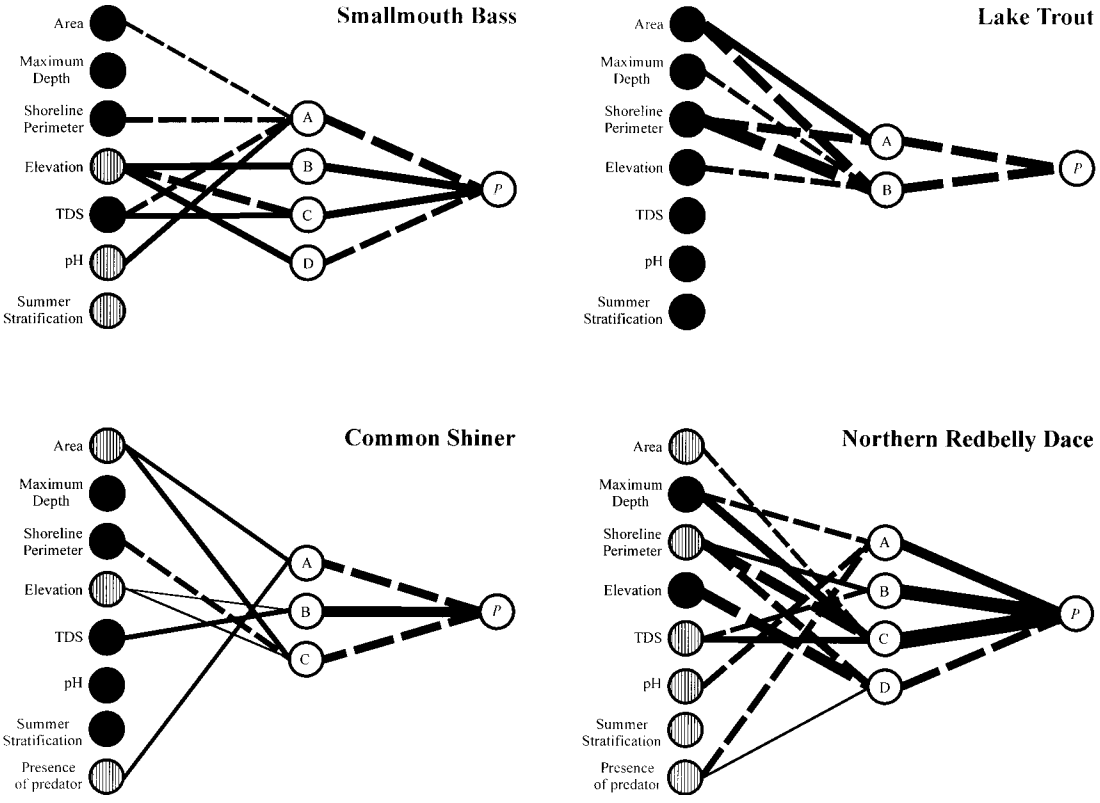

FIGURE 4.—Neural interpretation diagrams for predicting fish species presence/absence as a function of whole-lake habitat variables. The thickness of the lines joining neurons is proportional to the magnitude of the connection weight, and line type indicates the direction of the interaction between neurons: Solid line connections are positive (excitators) and dashed line connections are negative (inhibitors). All connection weights are statistically different from zero ($\alpha = 0.05$). Black input neurons indicate habitat variables that have an overall positive influence on species presence/absence; hatched input neurons indicate an overall negative influence on species presence/absence.

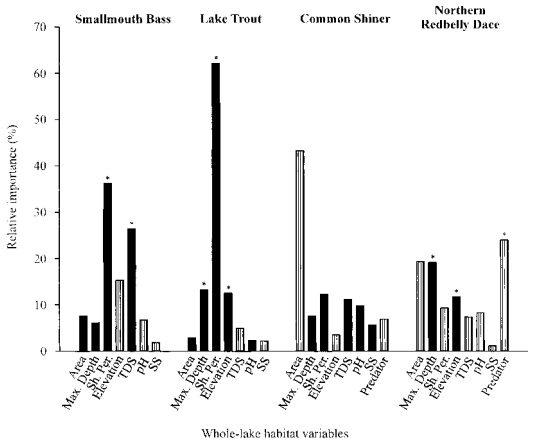

FIGURE 5.—The relative importance (% of total contribution) of whole-lake habitat variables in predicting the presence/absence of a species. Black bars indicate habitat variables that have an overall positive influence on species occurrence; hatched bars indicate an overall negative influence on species presence/absence. Asterisks indicate variables that make significant contributions according to the randomization test.

its input–hidden × hidden–output connection weights, and the relative contribution of each variable (expressed as a percent) is calculated by dividing the total contribution by the sum of the absolute total contributions of all variables in the network and multiplying by 100.

Individual and interacting influences of the habitat variables on the predicted probability of species occurrence were interpreted when connection weights differed significantly from random (based on $\alpha = 0.05$), and thus all neural interpretation diagrams illustrate only the connection weights that are nonrandom. The probability of smallmouth bass occurrence is positively correlated with lake area, shoreline perimeter, and TDS through hidden neuron A, as well as by lake elevation through hidden neuron B (Figure 4). In contrast, pH (hidden neuron A) and elevation (hidden neurons C and D) negatively influence the probability of occurrence. Focusing on hidden neuron C shows that the effects of lake elevation and TDS interact such that the negative influence of elevation on the probability of smallmouth bass occurrence weakens as TDS increases. Summing weights across all hidden neurons shows that shoreline perimeter and TDS have a significant positive effect on the predicted probability of smallmouth bass occurrence (Figure 5). The lake trout neural interpretation diagram shows that lake area and shoreline perimeter interact through hidden neuron A, resulting in the negative influence of surface area weakening as shoreline perimeter increases (Figure 4). Increasing maximum depth, shoreline perimeter, and elevation result in an increased probability of the occurrence of lake trout (Figure 5). Similar to lake trout, the probability of common shiner occurrence is affected by the interaction between area and shoreline perimeter (hidden neuron C; Figure 4). No habitat variables significantly contribute to predicted probabilities of common shiner occurrence, although lake area shows the strongest influence (Figure 5). The probability of northern redbelly dace occurrence decreases with the presence of a littoral-zone predator. However, this negative influence weakens with increasing shoreline perimeter and elevation (hidden neuron D; Figure 4). Maximum depth and elevation positively influence the probability of northern redbelly dace occurrence, whereas the presence of a littoral-zone predator has a strong negative influence (Figure 4).

The Madawaska lake models can be transferred readily to the Oxtongue drainage lakes, with rates of correct classification, sensitivity, and specificity being very similar for both drainages (Table 3). Because of differences in the frequency of occurrence of the species between the two drainages, only four of the nine species–habitat models differ significantly from random at the 5% level, although six of the nine are significant at a slightly less conservative level (i.e., $P \leq 0.063$). Most notably, common shiner, lake trout, and smallmouth bass are highly predictable in both the Madawaska and Oxtongue drainages. Moreover, for many species the optimal decision threshold for classifying a species as present or absent deviates from 0.5, but typically falls in the 0.4–0.6 range.

ANNs Versus Logistic Regression: Species Presence/Absence

We found that neural networks showed greater or equal correct classification rates than logistic models did for 14 of 18 fish–habitat models across the two drainages (Tables 3 and 4). On average ANNs outperformed logistic models by 2.2% for the Madawaska drainage and by 9.1% for the Oxtongue drainage; for many species, however, the difference between the approaches was greater. For example, common shiner, smallmouth bass, and northern redbelly dace were correctly predicted in an additional 16–41% of the Oxtongue lakes by using ANNs (Figure 6). Sensitivity and specificity values were similar for both approaches, although on average ANNs exhibited greater sensitivity and

TABLE 4.—Performance of logistic regression models in predicting species presence/absence in 128 lakes in the Madawaska River drainage (training data) based on *n*-fold cross validation and application of the Madawaska networks to predict presence/absence in 32 lakes from the Oxtongue River drainage (test data). The reported values are the percent species occurrence in the drainage (SO), the optimal decision threshold based on receiver-operating characteristic analysis (ODT) in which species with model probabilities greater than this value are predicted to be present, percent correct classification (CC), sensitivity (SE), specificity (SP), and the Kappa *z*-value statistic and associated *P*-value.

	Madawaska River drainage (training data)							Oxtongue River drainage (test data)					
Species	SO	ODT	CC	SE	SP	Kappa	*P*	SO	CC	SE	SP	Kappa	*P*
Brown bullhead	37.5	0.62	64.1	29.2	85.0	1.51	0.066	65.6	59.4	38.1	100	1.96	0.025
Common shiner	43.8	0.65	60.9	30.4	84.7	1.68	0.045	53.1	68.8	58.8	80.0	2.19	0.014
Creek chub	65.6	0.90	65.6	81.0	36.4	1.78	0.038	62.5	71.9	100	25.0	1.35	0.089
Golden shiner	35.2	0.49	64.8	8.9	95.2	0.43	0.334	37.5	65.6	8.3	100	0.46	0.325
Lake trout	43.0	0.46	77.3	65.5	86.3	5.74	0.000	56.3	75.0	88.9	57.1	2.56	0.005
Northern redbelly dace	53.1	0.52	55.5	77.9	30.0	0.90	0.184	28.1	31.3	100	4.4	0.22	0.414
Pumpkinseed	60.2	0.80	68.0	81.8	47.1	3.14	0.001	65.6	75.0	90.5	45.5	1.86	0.031
Smallmouth bass	25.0	0.39	75.0	0.0	100.0	0.00	0.500	46.9	53.1	0	100	0	0.500
Yellow perch	68.0	0.50	71.1	81.6	48.8	3.03	0.001	75.0	78.1	83.3	62.5	1.99	0.023
Mean	47.9		66.9	50.7	68.2	2.02		54.5	64.2	63.1	63.8	1.39	
SD	13.9		6.8	33.5	27.2	1.75		14.1	14.8	39.0	34.7	0.94	

logistic models exhibited greater specificity (Tables 3 and 4). A reduction in correct classification rates associated with the logistic models resulted in only five species–habitat models differing from random for the Madawaska lakes (compared with seven for ANNs). Five logistic models differed from random for the Oxtongue lakes (compared with four for ANNs), but this variability reflects a general lack of predictive directionality for the logistic models (i.e., more balanced levels of sensitivity and specificity).

ANNs for Fish Species Abundance

Within-lake variables predict species abundance with good accuracy and precision for creek chub ($r = 0.833$, RMSE = 0.194), golden shiner ($r = 0.783$, RMSE = 0.260), pumpkinseed ($r = 0.734$, RMSE = 0.209), and yellow perch ($r = 0.784$, RMSE = 0.204). The neural interpretation diagrams highlight relationships between predicted abundances and habitat for each species (Figure 7). For yellow perch the positive influence of wood cover on predicted abundance weakens with increasing density of vegetation (hidden neuron E), and the positive relationship between predicted abundance and depth diminishes with increasing site exposure (hidden neuron A; Figure 7). The amount of wood cover and depth contributes positively to the predicted yellow perch abundance, whereas vegetation density contributes negatively (Figure 8). Similarly, interactions among habitat variables for pumpkinseed abundance were common. The positive influence of wood cover and litter on predicted abundance weakens with increasing site exposure and depth (hidden neuron A; Figure 7). Accounting for all connection weights, increasing amounts of cover and litter and decreasing depth predict greater abundance of pumpkinseed (Figure 8). Predicted golden shiner abundance is negatively correlated with the amount of wood cover, but this relationship weakens with increasing depth (hidden neuron C; Figure 7). Overall, golden shiner abundance exhibits a positive association with vegetation density and a negative association with wood cover and sampling month (Figure 8). Predicted creek chub abundance is negatively associated with the presence of leaf litter and substrate type; however, this association diminishes with increasing depth (hidden neuron C; Figure 7). Vegetation density and depth have a positive influence, whereas leaf litter negatively influence predicted abundance of creek chub (Figure 8).

ANNs Versus Linear Regression: Species Abundance

We found that neural networks outperformed multiple linear regression models for predicting species abundance in Crosson Lake. Based on the regression model, within-lake variables predicted species abundance with moderate accuracy and precision for creek chub ($r = 0.810$, RMSE = 0.221), golden shiner ($r = 0.713$, RMSE = 0.298), pumpkinseed ($r = 0.689$, RMSE = 0.241), and yellow perch ($r = 0.731$, RMSE = 0.234). ANNs exhibited higher correlations between predicted

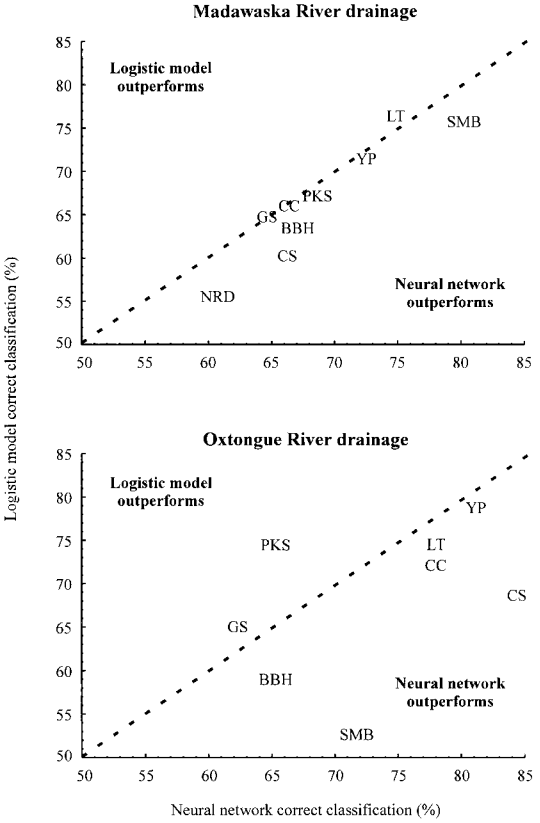

FIGURE 6.—Comparison of percent-correct classification rates for ANNs versus logistic regression models for presence/absence in the Madawaska and Oxtongue River drainages. Species codes refer to brown bullhead (BBH), creek chub (CC), common shiner (CS), golden shiner (GS), lake trout (LT), northern redbelly dace (NRD), pumpkinseed (PKS), smallmouth bass (SMB), and yellow perch (YP). Northern redbelly dace in the Oxtongue River drainage is not shown because the logistic regression model correctly classified its presence/absence in only 31.3% of the lakes, whereas the neural networks correctly classified its presence/absence in 71.9% of the lakes.

and observed abundances (ranging from 2.3% to 7.0%) and lower RMSE of predictions (ranging from 0.027 to 0.038) for all four species modeled.

Variable Selection in ANNs

In addition to using the results from the randomization test to interpret variable contributions, we used the approach as a variable selection method for removing input and hidden neurons for which incoming or outgoing connection weights were not significantly different from random. Retesting the predictive performance of these ''pruned'' networks, we found that the predictability of both species occurrence and abundance was generally unaffected by the removal of nonsignificant neurons in the network (Table 5). For example, the predictability of lake trout occurrence in the pruned network was similar to that in the original (unpruned) network, but the clarity of the network topology was improved. Ultimately, removing null hidden neurons and connection weights eases the interpretation of variable contributions in the network.

Discussion

Modeling Fish–Habitat Associations by Using ANNs

ANNs have several advantages over traditional modeling approaches that make them potentially beneficial for modeling fisheries data. ANNs are capable of modeling nonlinear associations for a variety of data types (e.g., continuous, discrete), require no specific assumptions concerning the distributional characteristics of the independent variables, and can accommodate interactions among predictor variables without any a priori specification (Ripley 1996). Because ANNs approximate any continuous function (Cybenko 1989; Funahashi 1989; Hornick et al. 1989), they exhibit flexibility for modeling nonlinear relationships between variables. For these reasons, the application of ANNs for pattern recognition and prediction has been advocated by researchers in several disciplines and has been shown in many ecological studies to exhibit greater predictive capabilities than traditional approaches such as regression-based techniques (e.g., Lek et al. 1996; Mastrorillo et al. 1997; Lek and Guégan 1999; this study). Indeed, the results from our study show that ANNs can provide a powerful quantitative approach for modeling fish–habitat relationships. Comparisons to traditional regression approaches showed that for almost all species ANNs provided greater power for predicting species occurrence and abundance. We stress, however, that where the underlying data structure and assumptions are met for a particular traditional statistical technique, there is no reason to believe that major differences will exist between traditional approaches and ANNs. The results of our simulation experiment support this perspective. ANNs were shown to be superior to regression approaches for nonlinearly distributed data (i.e., Gaussian species response curves), whereas both approaches showed identical predictive power for the logistic and linear species response curves. Nonetheless, given that eco-

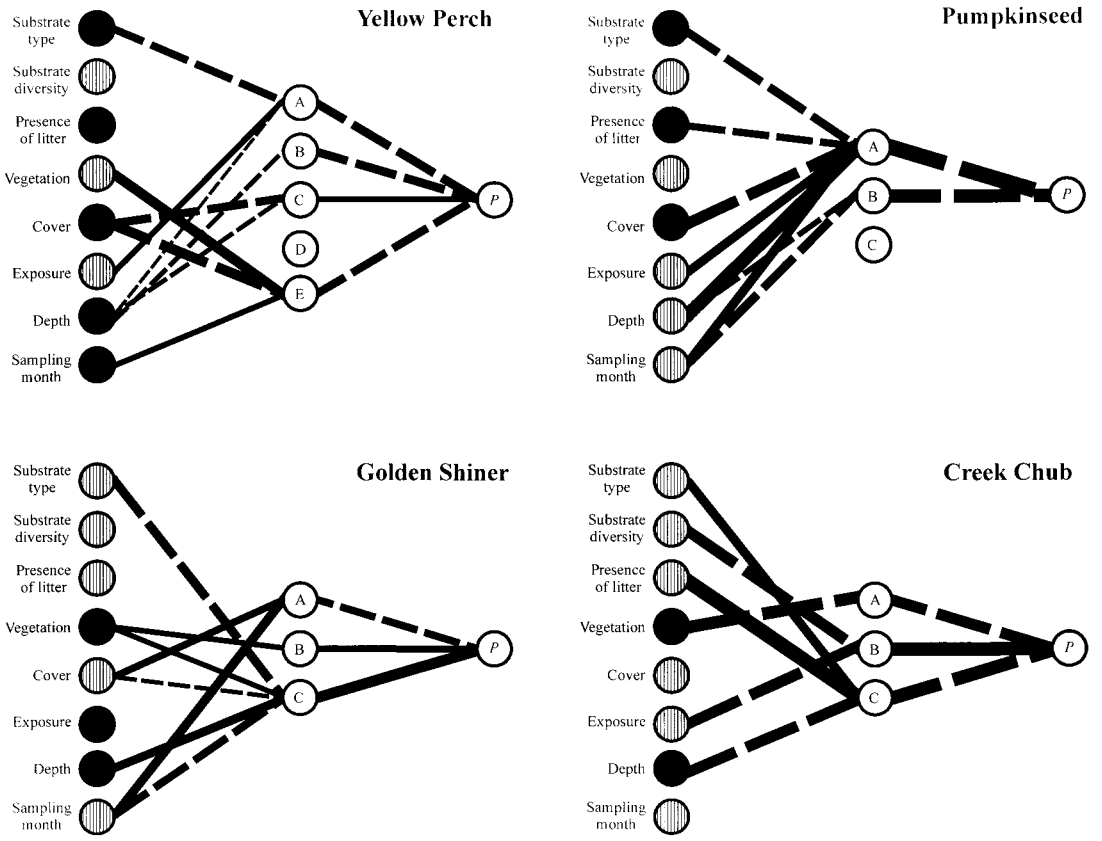

FIGURE 7.—Neural interpretation diagrams for predicting fish species abundance as a function of within-lake habitat variables. The thickness of the lines joining neurons is proportional to the magnitude of the connection weight, and line type indicates the direction of the interaction between neurons: Solid line connections are positive (excitators) and dashed line connections are negative (inhibitors). All connection weights are statistically different from zero ($\alpha = 0.05$). Black input neurons indicate habitat variables that have an overall positive influence on species abundance; hatched input neurons indicate an overall negative influence on species abundance.

logical data are commonly nonlinear in nature, that different solutions may arise due to specific choices of transformations, and achieving linearity is often not possible (e.g., Lek et al. 1996; Guégan et al. 1998; Wally and Fontama 1998), we believe ANNs provide an attractive alterative. Ultimately, more detailed simulation studies assessing similarities and differences between traditional and alternative statistical approaches using a broader variety of known data conditions are critically needed.

We have shown that species presence/absence was predictable from whole-lake measures of habitat, which is consistent with many studies of temperate fish populations (Jackson and Harvey 1989; Tonn et al. 1990; Magnuson et al. 1998). Species such as smallmouth bass and lake trout were predicted with high accuracy, an especially attractive result, given the economic and societal importance of these sport fishes. Similarly, ANNs provided accurate predictions of species abundance based on within-lake habitat characteristics. Although many researchers consider ANNs to have a practical disadvantage of failing to supply the explanatory insight provided by traditional approaches, our study shows that the contribution of the independent variables in the neural network can be quantified by direct evaluation of the connection weights. This examination is further aided by using a randomization approach to remove nonsignificant weights that do not contribute to the network prediction, thus assisting in the interpretation of direct and interacting effects of the variables in the network and simplifying the network structure (see Olden 2000b for another application of the randomization approach). For example, overall

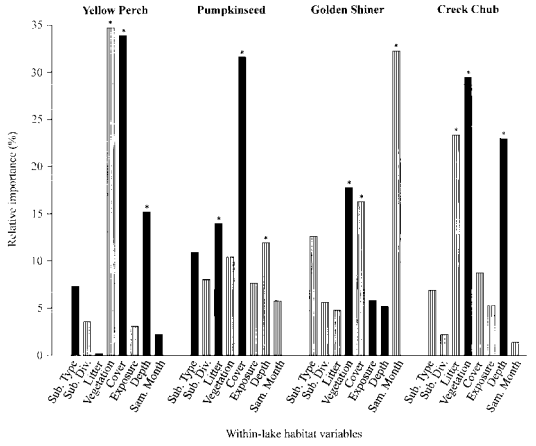

FIGURE 8.—Relative importance (% of total contribution) of within-lake habitat variables in predicting species abundance. Black bars indicate habitat variables that have an overall positive influence on species abundance; hatched bars indicate an overall negative influence on species abundance. Asterisks indicate variables that make significant contributions according to the randomization test.

lake size (i.e., area, maximum depth, and shoreline perimeter) and TDS (a surrogate for productivity) were identified as positively influencing the probability of smallmouth bass and lake trout occurrence. Lake area and maximum depth are known to influence the occurrence of these species (e.g., Eadie and Keast 1984; Jackson and Harvey 1989) because they alter the mixing characteristics and the thermal regime of lakes. Furthermore, lake area and depth serve as indirect measures of the diversity of habitats available in lakes, which may be important to support the small-bodied forage fish on which smallmouth bass and lake trout feed. Presence of a littoral-zone predator had a strong negative effect on the probability of northern redbelly dace occurrence but minimal effect on common shiner occurrence. This finding is consistent with studies that suggest the abundance and distributions of northern redbelly dace are greatly affected by the presence of a littoral predator (Findlay et al. 2000; MacRae and Jackson 2001), whereas common shiners appear to be more resistant to predation (Chapleau et al. 1997; Whittier et al. 1997). Interestingly, the negative relationship between northern redbelly dace and presence of a predator weakens substantially with increasing shoreline perimeter. As shoreline perimeter increases for a given lake area, the shoreline becomes more convoluted, increasing the potential for the presence of protected embayments and patchy nearshore habitats that provide increased habitat heterogeneity and potential refuge from predation (Jackson et al. 2001).

Several within-lake factors were related to increased species abundance. Greater abundances of yellow perch and pumpkinseed were predicted for sites with large amounts of coarse woody material and low densities of vegetation. The opposite was true for golden shiner and creek chub, which were found in greater abundance in more vegetated sites. Habitat cover was generally more important in the models for creek chub than those for golden shiner, supporting the view that creek chub populations may be less tolerant of habitat modifications (Whittier and Hughes 1998). Although the form of preferred cover differs among species, these results strengthen the notion that predicted

TABLE 5.—Comparison of model predictions of full and pruned neural networks. In the pruned networks, input variables and hidden neurons that were not statistically significant from zero (based on randomization test results) were removed; the pruned network design is given after the species name, with the three values representing the number of input, hidden, and output neurons, respectively. The reported values are the percent correct classification (CC), sensitivity (SE), and specificity (SP) for predicting species presence/absence (based on the optimal decision threshold from receiver-operating characteristic analysis) and the correlation coefficient (r) between predicted and actual abundances and the root mean square error of the prediction (RMSE) for predicting species abundance.

	Presence/absence						Abundance			
	Full network			Pruned network			Full network		Pruned network	
Species	CC	SE	SP	CC	SE	SP	r	RMSE	r	RMSE
Common shiner (5-3-1)	66.4	64.3	68.1	63.3	46.4	76.4				
Lake trout (4-2-1)	75.0	70.9	78.1	74.2	72.7	76.7				
Northern redbelly dace (7-4-1)	60.9	58.8	63.3	60.9	58.8	63.3				
Smallmouth bass (5-4-1)	80.5	90.6	50.0	79.7	95.8	31.3				
Creek chub (6-3-1)							0.833	0.194	0.813	0.206
Golden shiner (5-3-1)							0.783	0.260	0.748	0.281
Pumpkinseed (6-2-1)							0.734	0.209	0.622	0.236
Yellow perch (6-4-1)							0.784	0.204	0.779	0.206

abundance is greater in areas with greater habitat cover (Bryan and Scarnecchia 1992; Moring and Nicholson 1994; Christensen et al. 1996). Occupancy of complex habitats by golden shiner and creek chub supports the idea that these habitats provide profitable foraging areas (e.g., Werner et al. 1983; Diehl and Eklov 1995), rather than simply providing shelter from predation, because Crosson Lake lacks large piscivorous fish. Depth also played an important role in predicted species abundance. Yellow perch, creek chub, and golden shiner were predicted to be more abundant at depths of 1.5 m than at 0.5 m, whereas pumpkinseed was more numerous in shallower habitats closer to shore. Therefore, spatial occupancy of these species appears to be divided into different facets, depending on the interactions between the type of cover (i.e., vegetation or coarse woody material) and depth. In addition, these species–habitat associations were often influenced by the degree of site exposure. For example, the importance of depth and cover for predictions of yellow perch and pumpkinseed weakens with increasing site exposure. Finally, the fact that sampling month appears to be important for golden shiner abundance, which decreased from the July to the August sampling period, supports the importance of seasonal-dependent processes for some species in lakes (e.g., Hatzenbeler et al. 2000).

In summary, the ANNs provided a powerful technique for uncovering interactions among habitat characteristics of lakes and for determining their influence on species occurrence and abundance. Such interactions are more difficult to assess by multiple regression, which requires including multiplicative combinations of the variables directly into the models. For example, examining the direct and interactive effects of two variables, say, A and B, requires the inclusion of A, B, and A $\times$ B into the regression model. Where many variables are included, the number of possible variable combinations increases, contributing to increased type I errors in the results (see Olden and Jackson 2000 for details). ANNs do not require the inclusion of the additional interaction terms as separate variables.

Fish–Habitat Models as Important Management Tools

The development of models for predicting the distribution and abundance of fish populations is of paramount importance, given that demand continues for the development of lake shorelines. Our study shows that ANNs can provide accurate predictions regarding the abundance and occurrence of fish species based on within- and whole-lake habitat characteristics. Predictions about the effects of littoral-zone alteration on fish abundance could be a valuable tool for lake managers in deciding whether proposed shoreline modifications should be allowed in a system, or alternatively, deciding where in a lake the modifications should be permitted to minimize their impact on the fish community. Cottage owners often remove both macrophytes and woody material from their shorelines to enhance the cosmetic appearance of their property and minimize boating problems. Developed lakes with shoreline residences have substantially less density of coarse-woody material than do less developed lakes (Christensen et al. 1996), which can negatively affect species composition and fish abundances (e.g., Poe et al. 1986; Everett and Ruiz 1993) and decrease fish growth rates (e.g., Schindler et al. 2000). In addition, fish–habitat models may be particularly useful for predicting the cumulative effects of small-scale habitat modifications on fish abundance and spatial occupancy. Some researchers have argued that modeling the effects of small incremental habitat change may be impractical because of the difficulties in identifying and interpreting the effects of multiple modifications on fish populations (Jennings et al. 1999). Others have argued that modeling such relationships is not possible because of the lack of detailed data (Panek 1979) and of powerful quantitative techniques (Burns 1991). We believe that using alternative statistical methodologies help to offset these difficulties. Detailed data describing within-lake habitat characteristics and fish use currently exist for many systems. However, the most common approaches for analyzing and summarizing such data involve simple, descriptive statistics (Bain et al. 1999). Therefore, better use of available data and more flexible, powerful statistical methods, such as ANNs, may enable managers to predict the effects of small-scale habitat modifications on fish populations.

We have shown that whole-lake habitat attributes can successfully predict fish occurrence. The development of such models has important implications for prioritizing surveys and monitoring programs of fish populations because limits to resources preclude extensive sampling of aquatic habitats. Model predictions can also be used as first-order estimates of habitat suitability, which can be followed by ground-truthing and field validation, to predict sites with available spawning habitat (e.g., Knapp and Preisler 1999) or to es-

tablish potential locations for species reintroduction. Similarly, models can be used to predict the likelihood of local establishment and spread of exotic species, which may help set conservation priorities for preserving vulnerable species and populations that might be lost locally (e.g., Hrabik and Magnuson 1999).

Enhancing the Predictive Power of Fish–Habitat Models

The predictive abilities of conventional models for species presence/absence are commonly assessed from overall classification rates alone. We show that by partitioning the predictive performance of the models into measures such as sensitivity and specificity, we can more readily assess the strengths and weaknesses of the models. For example, the presence of creek chub, pumpkinseed, and yellow perch could be predicted with a high degree of certainty (in more than 90% of the lakes); predicting the absence of these species, however, was more difficult. It is also evident that model sensitivity increases and specificity decreases with increasing frequency of species occurrence (i.e., species prevalence) in the lakes. This relationship is expected yet is seldom considered in distribution modeling. The relationship between prediction success and species prevalence in the data set has several practical implications. First, a decrease in model sensitivity for rare species implies it will be more difficult to predict the occurrence of organisms for which conservation and management are most critical. Consequently, our ability to identify suitable locations for species reintroductions could be limited. Second, drawing inferences from observed absences of species from sites containing suitable habitat conditions (e.g., indirect evident for dispersal, predation, and competition) could be limited if the models show poor specificity. Examining alternative measures of prediction success can provide more accurate comparisons of different modeling approaches (e.g., Manel et al. 1999) and different models (i.e., different subsets of variables). For example, we found that although the overall correct classification rates for some species were similar, specificity and sensitivity values were often quite different. Also, correct classification rates did not change between the full and the pruned neural networks, but sensitivity and specificity both did.

The effect of species prevalence in model development is unavoidable; one would expect that, given an increased frequency of occurrence, the probability of predicting the species to be present is greater. However, varying the decision threshold probability for which the model predicts presence/absence, rather than following the conventional arbitrary threshold of 0.5, can compensate for this bias and result in more powerful models (e.g., Carroll et al. 1999; Manel et al. 1999). Determining the optimal decision threshold involves constructing ROC plots and then choosing the threshold that maximizes sensitivity and specificity, given particular misclassification costs. This technique has been applied widely to clinical problems in medicine, but few ecological studies have used ROC analysis. We used equal costs of false presence (misclassifying a species as present) and false absence (misclassifying a species as absent); however, in practice, it may be advantageous to assign more appropriate costs to the misclassifications if such information is available. Although assigning costs is a complex and potentially subjective process, much can be gained. For example, we might tolerate more false presences for endangered species rather than fail to protect a species and thus could adjust the decision threshold accordingly to develop a more powerful predictive model.

Finally, one important concern is that many models lack geographical transferability (i.e., poor model performance outside the original data used to develop the model) because species–environment associations can differ substantially in different systems (e.g., Leftwich et al. 1997). Nevertheless, models may be useful when applied at the scale from which they were developed and in systems where similar species–environment associations exist. We have shown that testing models in adjacent drainages demonstrates the generality of the fish–habitat models. Models built using lakes in the Madawaska River drainage not only performed well for the same set of lakes, but actually performed slightly better, on average, for predicting species occurrence in the Oxtongue River drainage. Although one might be surprised that correct classification rates were slightly higher in the Oxtongue lakes (i.e., test data) compared with the Madawaska lakes (i.e., training data), in fact model sensitivity was on average high and the species modeled were more prevalent in the Oxtongue lakes than in the Madawaska drainage. Consequently, the effect of species prevalence on geographic transferability of fish–habitat models also needs to be considered.

Conclusion

ANNs have wide applicability to the study of ecological relationships, both as exploratory and

predictive tools. ANNs provide a flexible approach that can accommodate a wide variety of study designs without the statistical constraints of independence and linearity, and they require no a priori understanding of variable relationships. Consequently, they are useful techniques for relating the distributions and abundances of fish populations to their physical environment. Given the obvious importance of establishing linkages between habitat features, fish distributions, and the use of nearshore habitats by fish, the development and testing of fish–habitat models are important steps in the conservation and management of lake fish populations. Such predictive models can advance management efforts to understand fish–habitat associations and predict the effects of natural and anthropogenic-related habitat modification on freshwater fish populations.

Acknowledgments

We thank Nick Collins, Scott Hinch, Nigel Lester, Brian Shuter, Keith Somers, and two anonymous reviewers for their constructive comments on various versions of this paper. Special thanks to Nick Mandrak for providing the Algonquin Park data and for numerous enjoyable discussions regarding the fish communities of Ontario. We are also indebted to Sovan Lek for discussing the finer statistical details of neural networks. Funding for this research was provided by a Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and University of Toronto scholarships to J.D.O., and an NSERC Research Grant to D.A.J.

References

Allen, J. D., and A. S. Flecker. 1993. Biodiversity conservation in running waters. Bioscience 43:32–43.

Aoki, I., and T. Komatsu. 1997. Analysis and prediction of the fluctuations of sardine abundance using a neural network. Oceanologica Acta 20:81–88.

Bain, M. B., T. C. Hughes, and K. K. Arend. 1999. Trends in methods for assessing freshwater habitats. Fisheries 24:16–21.

Beauchamp, D. A., E. R. Byron, and W. A. Wurtsbaugh. 1994. Summer habitat use by littoral-zone fishes in Lake Tahoe and the effects of shoreline structures. North American Journal of Fisheries Management 14:385–394.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Brosse, S., J. F. Guégan, J. N. Tourenq, and S. Lek. 1999. The use of neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120:299–311.

Bryan, M. D., and D. L. Scarnecchia. 1992. Species richness, composition, and abundance of fish larvae and juveniles inhabiting natural and developed shorelines of a glacial Iowa lake. Environmental Biology of Fishes 35:329–341.

Burns, D. C. 1991. Cumulative effects of small modifications to habitat. Fisheries 16(1):12–17.

Carroll, C., W. J. Zielinski, and R. F. Noss. 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.

Chapleau, F., C. S. Findlay, and E. Szenasy. 1997. Impact of piscivorous fish introductions on fish species richness of small lakes in Gatineau Park, Quebec. Écoscience 4:259–268.

Chen, D. G., and D. M. Ware. 1999. A neural network model for forecasting fish stock recruitment. Canadian Journal of Fisheries and Aquatic Sciences 56:2385–2396.

Christensen, D. L., B. J. Herwig, D. E. Schindler, and S. R. Carpenter. 1996. Impacts of lakeshore residential development on course woody debris in north temperate lakes. Ecological Applications 6: 1143–1149.

Colasanti, R. L. 1991. Discussions of the possible use of neural network algorithms in ecological modelling. Binary 3:13–15.

Crossman, E. J., and N. E. Mandrak. 1991. An analysis of fish distribution and community structure in Algonquin Park: annual report for 1991 and completion report, 1989–1991. Ontario Ministry of Natural Resources, Toronto, Ontario, Canada.

Cybenko, G. 1989. Approximation by superimpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2:303–314.

Dodge, D. P., G. A. Goodchild, I. MacRitchie, J. C. Tilt, and D. G. Waldriff. 1985. Manual of instructions: aquatic habitat inventory surveys. Ontario Ministry of Natural Resources, Fisheries Branch, Toronto, Ontario, Canada.

Diehl, S., and P. Eklov. 1995. Piscivore-mediated habitat use in fish: effects on invertebrate resources, diet, and growth of perch, *Perca fluviatilus*. Ecology 76: 1712–1726.

Dimopoulos, Y., P. Bourret, and S. Lek. 1995. Use of some sensitivity criteria for choosing networks with good generalization. Neural Processing Letters 2: 1–4.

Eadie, J. M., and A. Keast. 1984. Resource heterogeneity and fish species diversity in lakes. Canadian Journal of Zoology 62:1689–1695.

Edwards, M., and D. R. Morse. 1995. The potential for computer-aided identification in biodiversity research. Trends in Ecology and Evolution 10:153–158.

Everett, R. A., and G. M. Ruiz. 1993. Coarse woody debris as a refuge from predation in aquatic communities: an experimental test. Oecologia 93:475–486.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.

Findlay, C. S., D. G. Bert, and L. Zheng. 2000. Effect of introduced piscivores on native minnow communities in Adirondack lakes. Canadian Journal of Fisheries and Aquatic Sciences 57:570–580.

Funahashi, K. 1989. On the approximate realization of continuous mapping by neural networks. Neural Networks 2:183–192.

Garson, G. D. 1991. Interpreting neural-network connection weights. Artificial Intelligence Expert 6:47–51.

Geman, S., E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. Neural Computation 4:1–58.

Guégan, J. F., S. Lek, and T. Oberdorff. 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature (London) 391:382–384.

Harig, A. L., and M. B. Bain. 1998. Defining and restoring biological integrity in wilderness lakes. Ecological Applications 8:71–87.

Hatzenbeler, G. R., M. A. Bozek, M. J. Jennings, and E. E. Emmons. 2000. Seasonal variation in fish assemblage structure, and habitat structure in the nearshore littoral zone of Wisconsin Lakes. North American Journal of Fisheries Management 20:360–368.

Hinch, S. G., N. C. Collins, and H. H. Harvey. 1991. Relative abundance of littoral zone fishes: Biotic interactions, abiotic factors, and postglacial colonization. Ecology 72:1314–1324.

Hornick, K., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366.

Hrabik, T. R., and J. J. Magnuson. 1999. Simulated dispersal of exotic rainbow smelt (*Osmerus mordax*) in a northern Wisconsin lake district and implications for management. Canadian Journal of Fisheries and Aquatic Sciences 56(Suppl. 1):35–42.

Hughes, R. M., and R. F. Noss. 1992. Biological diversity and biological integrity: current concerns for lakes and streams. Fisheries 17(3):11–19.

Jackson, D. A., and H. H. Harvey. 1989. Biogeographic associations in fish assemblages: local versus regional processes. Ecology 70:1472–1484.

Jackson, D. A., and K. M. Somers. 1989. Are probability estimates from the permutation model of Mantel's test stable? Canadian Journal of Zoology 67:766–769.

Jackson, D. A., and H. H. Harvey. 1997. Qualitative and quantitative sampling of lake fish communities. Canadian Journal of Fisheries and Aquatic Sciences 54:2807–2813.

Jackson, D. A., P. R. Peres-Neto, and J. D. Olden. 2001. What controls who is where in freshwater fish communities?—the roles of biotic, abiotic, and spatial factors. Canadian Journal of Fisheries, and Aquatic Sciences 58:157–170.

James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Reviews in Ecology and Systematics 21:129–166.

Jennings, M. J., K. Johnson, and M. Staggs. 1996. Shoreline protection study: a report to the Wisconsin state legislature. Wisconsin Department of Natural Resources, Publication PUBL-RS-921–96, Madison.

Jennings, M. J., M. A. Bozek, G. R. Hatzenbeler, E. E. Emmons, and M. D. Staggs. 1999. Cumulative effects of incremental shoreline habitat modification on fish assemblages in north temperate lakes. North American Journal of Fisheries Management 19:18–27.

Knapp, R. A., and H. K. Preisler. 1999. Is it possible to predict habitat use by spawning salmonids? A test using California golden trout (*Oncorhynchus mykiss aguabonita*). Canadian Journal of Fisheries and Aquatic Sciences 56:1576–1584.

Kurková, V. 1992. Kolmogorov's theorem and multilayer neural networks. Neural Networks 5:501–506.

Leftwich, K. N., P. L. Angermeier, and C. A. Dolloff. 1997. Factors influencing behaviour and transferability of habitat models for a benthic stream fish. Transactions of the American Fisheries Society 126: 725–734.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90:39–52.

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

Lester, N. P., W. I. Dunlop, and C. C. Willox. 1996. Detecting changes in the nearshore fish community. Canadian Journal of Fisheries and Aquatic Sciences 53(Suppl. 1):391–402.

MacRae, P. S. D., and D. A. Jackson. 2001. The influence of predation, and habitat complexity on the structure of littoral-zone fish assemblages. Canadian Journal of Fisheries and Aquatic Sciences 58:342–351.

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

Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36:734–747.

Mastrorillo, S., S. Lek, F. Dauba, and A. Beland. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biology 38:237–246.

Matuszek, J. E., and G. L. Beggs. 1988. Fish species richness in relation to lake area, pH, and other abiotic factors in Ontario lakes. Canadian Journal of Fisheries and Aquatic Sciences 45:1931–1941.

Metz, C. E. 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8:283–298.

Minns, C. K. 1989. Factors affecting fish species richness in Ontario lakes. Transactions of the American Fisheries Society 118:533–545.

Moring, J. R., and P. H. Nicholson. 1994. Evaluation of three types of artificial habitats for fishes in a freshwater pond in Maine, USA. Bulletin of Marine Science 55:1149–1159.

Olden, J. D. 2000a. Predictive models for freshwater fish community composition. Masters thesis. University of Toronto, Ontario, Canada.

Olden, J. D. 2000b. An artificial neural network approach to studying phytoplankton succession. Hydrobiologia 436:131–143.

Olden, J. D., and D. A. Jackson. 2000. Torturing data for the sake of generality: How valid are our regression models? Écoscience 7:501–510.

Olden, J. D., D. A. Jackson, and P. R. Peres-Neto. 2001. Spatial isolation in freshwater drainage lakes. Oecologia 127:572–585.

Özesmi, S. L., and U. Özesmi. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecological Modelling 116: 15–31.

Panek, F. M. 1979. Cumulative effects of small modifications to habitat. Fisheries 4(2):54–57.

Poe, T. P., C. O. Hatcher, C. L. Brown, and S. W. Schloesser. 1986. Comparison of species composition and richness of fish assemblages in altered and unaltered littoral habitats. Journal of Freshwater Biology 3: 525–536.

Richter, B. D., D. P. Braun, M. A. Mendelson, and L. L. Master. 1997. Threats to imperiled freshwater fauna. Conservation Biology 11:1081–1093.

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

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagation errors. Nature (London) 323:533–536.

Schindler, D. E., S. I. Geib, and M. R. Williams. 2000. Patterns in fish growth along a residential development gradient in north temperate lakes. Ecosystems 3:229–237.

Titus, K., J. A. Mosher, and B. K. Williams. 1984. Chance-corrected classification for use in discriminant analysis: Ecological applications. American Midland Naturalist 111:1–7.

Tonn, W. M., J. J. Magnuson, M. Rask, and J. Toivonen. 1990. Intercontinental comparison of small-lake fish assemblages: The balance between local and regional processes. American Naturalist 136:345–375.

Walley, W. J., and V. N. Fontama. 1998. Neural network predictors of average score per taxon and number of families at unpolluted sites in Great Britain. Water Resources 32:613–622.

Werner, E. E., G. G. Mittelbach, D. J. Hall, and J. F. Gilliam. 1983. Experimental tests of optimal habitat use in fish: the role of relative habitat profitability. Ecology 64:1525–1539.

Whittier, T. R., D. B. Halliwell, and S. G. Paulsen. 1997. Cyprinid distributions in Northeast U.S.A. lakes: evidence of regional-scale minnow biodiversity losses. Canadian Journal of Fisheries and Aquatic Sciences 54:1593–1607.

Whittier, T. R., and R. M. Hughes. 1998. Evaluation of fish species tolerances to environmental stressors in lakes in the northeastern United States. North American Journal of Fisheries Management 18:236–252.

Williams, J. E., J. E. Johnson, D. A. Hendrickson, S. Contreras-Balderas, J. D. Williams, M. Navarro-Mendoza, D. E. McAllister, and J. E. Deacon. 1989. Fishes of North America: endangered, threatened, or of special concern. Fisheries 14(6):2–20.