Predictive Models of Fish Species Distributions: A Note on Proper Validation and Chance Predictions

JULIAN D. OLDEN,*[1] DONALD A. JACKSON, AND PEDRO R. PERES-NETO

Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario M5S 3G5,
Canada

Abstract.—The prediction of species distributions is a primary goal in the study, conservation, and management of fisheries resources. Statistical models relating patterns of species presence or absence to multiscale habitat variables play an important role in this regard. Researchers, however, have paid little attention to how improper model validation and chance predictions can result in unfounded confidence in the performance and utility of such models. Using simulated and empirical data for 40 lake and stream fish species, we demonstrate that the commonly employed resubstitution approach to model validation (in which the same data are used for both model construction and prediction) produces highly biased estimates of correct classification rates and consequently an inaccurate perception of true model performance. In contrast, a jackknife approach to validation resulted in relatively unbiased estimates of model performance. The estimated rates of model correct classification are also shown to be substantially influenced by species prevalence (i.e., the proportion of sites at which a species is present), and often result in poorly performing models being viewed as powerful. We use simulated data to show how the expected frequency of chance predictions from models is a function of species prevalence and sample size. Finally, we use empirical data to illustrate a randomization approach for assessing whether the performances of the fish habitat models are statistically greater than expectations based on chance predictions. In summary, we urge researchers to employ proper and defensible methodologies for model validation and prediction assessment; failing to do so will only add to the accumulating number of published species habitat models in the fisheries literature that are of limited use and reliability.

Knowledge of the current and future status of the biological diversity and integrity of lakes and streams is of paramount importance (Hughes and Noss 1992; Harig and Bain 1998; Hawkins et al. 2000). Recent attention has focused on the modification and loss of aquatic habitats as primary factors threatening the conservation of fish populations and communities (Williams et al. 1989;

Richter et al. 1997). Consequently, efforts to understand the link between local and regional habitat factors and habitat use by fish have become increasingly important and today represent a central avenue of research (Jackson et al. 2001). Statistical models play an important role in this regard not only by contributing to our understanding of fish habitat relationships but also by providing a framework from which spatial and temporal patterns in species presence or absence can be predicted. Consequently, species habitat models can contribute significantly to the management and conservation of fish populations and communities. For example, model predictions can be used to (1) estimate habitat suitability, (2) forecast the effects of habitat change due to altered land-use patterns (Oberdorff et al. 2001) and climate warming (Stefan et al. 2001), (3) establish potential locations for species reintroduction (Evans and Olver 1995), (4) predict the likelihood of invasion of exotic species (Peterson and Vieglais 2001), and (5) predict ''hotspots'' of species persistence for the conservation of biodiversity (Williams and Araujo 2000).

Although numerous modeling approaches are available for quantifying species habitat relationships (see Guisan and Zimmermann 2000 for a recent review), there is a generally held notion among researchers that once a model has been constructed and predictions have been produced (regardless of the statistical approach used), the modeling process is complete. This view is further reenforced by the automated nature of statistical modeling procedures in software packages, but, unfortunately, it is neither accurate nor suitable. Rather, proper model validation and the assessment of model predictions are two critical issues that must be addressed in order to accurately evaluate model performance and future applicability. Model validation refers to the application of a model, the generation of predictions, and the quantification of predictive performance (e.g., the correct classification rate for predicting species presence or absence). Validation is an important step in the modeling process because it quantifies our confidence in the predictions produced from future

---

applications of the model. After proper model validation, it is critical to assess whether the predictions produced by the model differ from what would be expected based on chance alone. This assessment is another important step in the modeling process because models producing random predictions are neither helpful nor useful. Although the validation and assessment of model predictions are two key considerations when developing species habitat models, these issues receive little or no attention by fisheries researchers and researchers in general (see Manel et al. 2001), and therefore the seriousness of their consequences are not appreciated. To address this concern, we use both simulated and empirical data to demonstrate the potential problems (and subsequent implications) of improper validation and chance predictions as they effect the development and interpretation of species presence or absence models. We hope that by presenting this study and highlighting the magnitude and consequences of these problems, we can help ensure that fisheries biologists and managers avoid them in the future.

Model Validation

Although logic dictates that a model has little merit if its predictions are not accurately validated with independent data, aquatic researchers commonly ignore this consideration in their research (Verbyla and Litaitis 1989; Olden and Jackson 2000). The traditional approach to model validation uses the same data both to construct the model and to estimate its predictive performance (known as the resubstitution estimate of the correct classification rate; Fielding and Bell 1997). Statisticians have long recognized that this approach leads to a biased estimate of model performance because the model has been optimized to deal with the unique characteristics and ''noise'' in the data set, and therefore has lost predictive generality beyond the original data (e.g., Efron 1986; Chatfield 1995). To obtain a more realistic estimate of model performance, one must validate the model with data that are independent of the data used to estimate the model's parameters. Cross-validation is a class of approaches that accomplish this goal and thus provides a more appropriate method for estimating the correct classification rates (i.e., the correct prediction of species presence or absence) of statistical models. Two types of cross-validation approaches, data splitting and jackknife validation, have been shown to be less biased than a resubstitution approach (Efron 1986; Olden and Jackson 2000). The jackknife method excludes one observation, constructs the model with the remaining $n - 1$ observations, and then predicts the response of the excluded observation using this model. The procedure is repeated n times so that each observation, in turn, is excluded in model construction and its response is predicted.

Given the importance of obtaining an accurate representation of model predictive performance, it is surprising that researchers predominantly use the resubstitution approach to validate their models (Manel et al. 2001). Motivated by this observation, we seek to demonstrate the extent to which resubstitution validation produces biased estimates of model predictive performance. We illustrate this bias by developing logistic regression models using simulated and empirical data for fish species occurrence in the temperate lakes and streams of Canada.

Simulated Data

Eleven prevalence classes of data were examined. The level of species prevalence (i.e., the proportion of sites at which a species is present) differed among classes and ranged from 0% to 100% (with incremental increases of 10%); random associations between the dependent and independent variables were maintained. For each prevalence class we simulated 10,000 data sets, each containing 50 observations of one dependent and four independent variables (all variables exhibited random, or uncorrelated, relationships with each other). Independent variables were generated by random sampling from a uniform distribution, although comparable results were obtained using normally distributed random data. The following Monte Carlo experiment was then repeated for each of the 10,000 data sets: (1) logistic regression analysis was conducted to model the relationship between the dependent and independent data; (2) the resubstitution and jackknife estimate of correct classification (%) was calculated (i.e., the percent of observations, either present or absent, that were correctly predicted); and (3) the difference between the resubstitution and jackknife correct classification rates and the ''random'' classification rate (related to species prevalence; see Chance Predictions below) was calculated. The mean difference based on the 10,000 data sets for each prevalence class provides a quantitative measure of the bias associated with resubstitution and jackknife model validation approaches because the true relationship between the dependent and independent variables is known (i.e., the correct classification rates expected from random data). Figure 1 shows
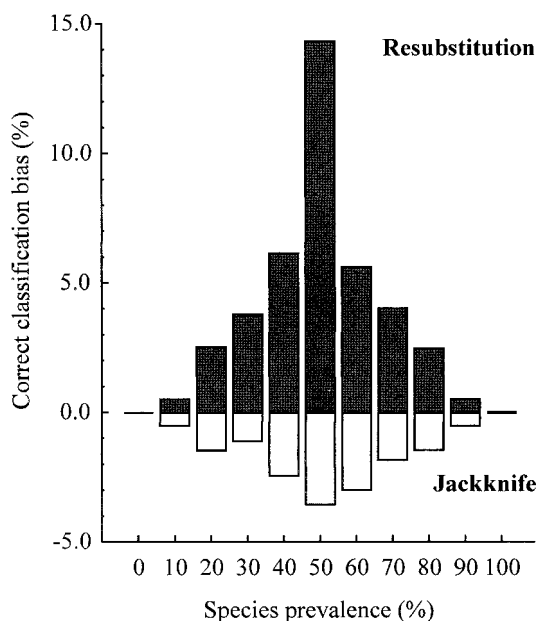
FIGURE 1.—Results from the Monte Carlo experiment showing biases in the resubstitution and jackknife estimates of correct classification (%). The term ''bias'' refers to the difference between estimated correct classification rates and estimates based on data exhibiting random associations between the dependent and independent variables. Mean differences are based on 10,000 simulated data sets and are reported for different classes of species prevalence.

the overestimation of the correct classification rates based on the resubstitution method, with biases approaching 15% (i.e., the resubstitution model reports that it can correctly classify 15% more of the observations than is truly the case). Note that the degree of bias associated with the resubstitution method increases from either direction as species prevalence approaches 50%. In contrast, the jackknife validation approach never overestimates the correct classification rate and, in fact, produced slight underestimations (with a maximum of less than 4%). This finding agrees with the findings of other simulation studies that have shown the jackknife approach to be a relatively unbiased approach (Olden and Jackson 2000).

Empirical Data

We constructed fish habitat models using logistic regression analysis for two data sets—one involving 89 lakes in the Petawawa River basin of Ontario, Canada, and one involving 48 streams in the Blackwater River basin of British Columbia, Canada. The Petawawa River data set consisted of 26 fish species (hereafter called the Petawawa species) and the following six whole-lake habitat variables: lake area, maximum depth, shoreline perimeter, lake elevation, total dissolved solids, and pH (see Crossman and Mandrak 1992; Olden 2000). The Blackwater River data set consisted of 14 species (hereafter called the Blackwater species) and the following four basin-scale habitat variables: drainage area, watershed gradient, reach gradient, and elevation (see Porter et al. 2000). We assessed correct classification rates for both data sets using the resubstitution and jackknife methods.

Figure 2 shows the jackknife estimates versus the resubstitution estimates of correct classification rate based on the fish habitat models for the Petawawa and Blackwater species. The vertical length of the residual from each point (representing a species) to the diagonal 1:1 line indicates the magnitude of the difference between the resubstitution and the jackknife estimate of correct classification and, therefore, the magnitude of bias in the resubstitution approach relative to the more accurate jackknife estimates. We use the term ''bias'' to describe the known overestimation of the resubstitution correct classification rate compared with the jackknife estimate based on the previous simulation experiments. For the Petawawa models, correct classification rates were inflated by over 5% on average (ranging from 1.1% to 12.3%), with pearl dace (12.3%), yellow perch (9%) and northern redbelly dace (9%) showing the greatest biases (see Figure 2 for data and scientific names). Similarly, average correct classification rates for the Blackwater models were inflated by 6% (ranging from 0% to 10.5%) based on the resubstitution approach, where the greatest differences (i.e., correct classification biases exceeding 10%) were observed for bridgelip sucker, leopard dace, and longnose dace (Figure 2). The insets of Figure 2 illustrate the frequencies of resubstitution bias for each data set and show, for example, that the majority of Petawawa models exhibited a correct classification bias between 3% and 7%, whereas the bias associated with the Blackwater models ranged uniformly between 1% and 10%. The two empirical examples show that the resubstitution estimates of fish habitat model performance would be interpreted to be substantially better than they really are, ultimately resulting in undue confidence in the overall quality and applicability of the models. In a similar study of 52 lakes, we found that the resubstitution estimate of the correlation coefficient between the predicted and actual values
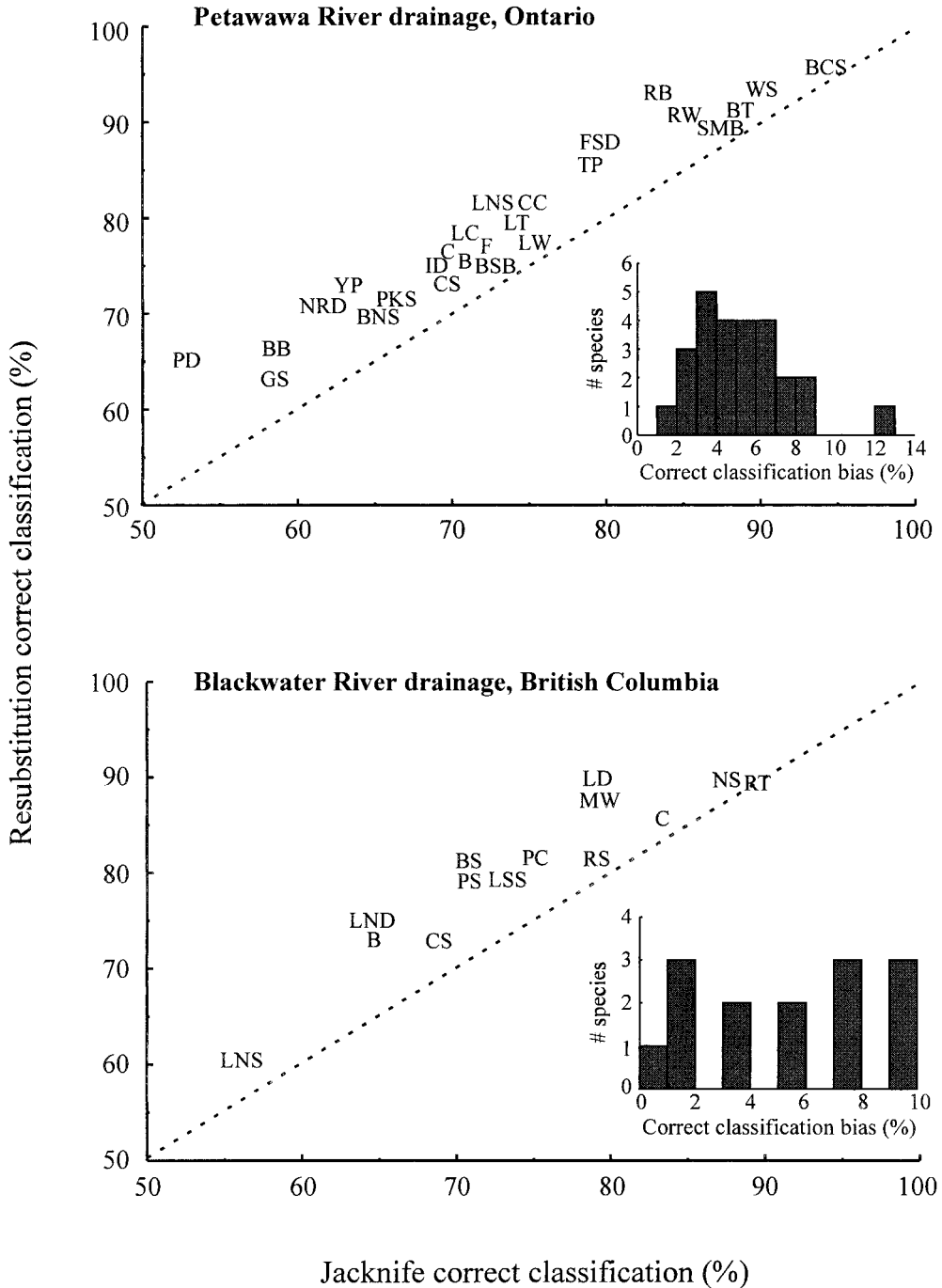
FIGURE 2.—The relationship between jackknife and resubstitution estimates of correct classification for 26 fish species in Petawawa River drainage, Ontario, and 14 fish species in Blackwater River drainage, British Columbia. Petawawa species codes refer to burbot *Lota lota* (B), brown bullhead *Ameiurus nebulosus* (BB), blackchin shiner *Notropis heterodon* (BCS), blacknose shiner *N. heterolepis* (BNS), brook stickleback *Culaea inconstans* (BSB), brook trout *Salvelinus fontinalis* (BT), cisco *Coregonus artedi* (C), creek chub *Semotilus atromaculatus* (CC), common shiner *Luxilis cornutus* (CS), fallfish *Semotilus corporalis* (F), finescale dace *Phoxinus neogaeus* (FSD), golden shiner *Notemigonus crysoleucas* (GS), Iowa darter *Etheostoma exile* (ID), lake chub *Couesius plumbeus* (LC), longnose sucker *Catostomus catostomus* (LNS), lake trout *Salvelinus namaycush* (LT), lake whitefish *Coregonus*

of species richness was 0.77, a figure considerably greater than the more accurate level of 0.52 that was estimated using a jackknife approach (Olden and Jackson 2000).

Chance Predictions

Are model predictions better than those that can be obtained by chance? A simple answer might involve generating chance predictions by flipping a coin to decide whether a species is predicted to be present or absent at a site, resulting in a 50% correct classification rate. However, this naïve approach is inappropriate because chance predictions are actually related to species prevalence (i.e., the frequency of species occurrence in the data set). For example, if one wished to predict the occurrence of a species at a particular site and the species was present in 80% of the sites surveyed, the probability of chance correct classification is not 50%, but 80%. That is, under the null hypothesis that a predictive model performs no better than random assignment, the number of correctly classified cases approximates a binomial distribution, with chance correct classification probability equal to $0.5 + |(0.5 -$ species prevalence)$|$. This results in a V-shaped response of percent correct classification to increasing levels of species prevalence. As an example, a trivial model for a rare species (e.g., 5% prevalence in the data set) would predict the species to be absent from all sites, resulting in a 95% correct classification rate. Similarly, a common species found at 95% of the sites would be predicted to occur at all sites, and, again, the model would correctly classify 95% of the sites. Finally, using such a trivial model for a species occurring at 50% of the sites would result in the maximum error rate, producing correct classifications for 50% of the sites. In this study we use simulated and empirical data to show how the relationship between chance predictions and species prevalence can erroneously enhance our perception of model predictive performance.

Simulated Data

Nineteen prevalence classes of data were examined. The level of species prevalence differed among classes and ranged from 5% to 95% (with incremental increases of 5%); random associations between the dependent and independent variables were maintained. For each prevalence class we simulated 10,000 data sets for each of three sample sizes (30, 60, and 90 observations) of one dependent and four independent variables (all variables exhibited random, or uncorrelated relationships with each other). Independent variables were generated by random sampling from a uniform distribution, although comparable results were obtained using normally distributed random data. The following Monte Carlo experiment was then repeated for each of the 10,000 data sets for each sample size: (1) logistic regression analysis was conducted to model the relationship between the dependent and independent data; and (2) the jackknife estimate of the correct classification rate (%) was calculated. Based on the 10,000 data sets for each prevalence class and for each sample size, 95% confidence intervals of the jackknife estimates of percent correct classification were calculated. Figure 3 shows the 95% confidence intervals for the three sample sizes across the entire spectrum of species prevalence and illustrates the expected range of correct classification rates from a model built using completely random data. It is important to note that mean correct classification rates due to chance show close agreement with the formula defined previously, and that the range of correct classification rates decreases with increasing sample size as well as with movement away from a species prevalence of 50% (Figure 3). Results from the simulation study highlight the over-

---

←

clupeaformis (LW), northern redbelly dace *Phoxinus eos* (NRD), pearl dace *Margariscus margarita* (PD), pumpkinseed *Lepomis gibbosus* (PKS), rock bass *Ambloplites rupestris* (RB), round whitefish *Prosopium cylindraceum* (RW), smallmouth bass *Micropterus dolomieu* (SMB), trout-perch *Percopsis omiscomaycus* (TP), white sucker *Catostomus commersoni* (WS) and yellow perch *Perca flavescens* (YP). Blackwater species codes refer to bridgelip sucker *Catostomus columbianus* (BS), burbot (B), chinook salmon *Oncorhynchus tshawytscha* (CS), chiselmouth *Acrocheilus alutaceus* (C), largescale sucker *Catostomus machrocheilus* (LSS), leopard dace *Rhinichthys falcatus* (LD), longnose dace *R. cataractae* (LND), longnose sucker (LNS), mountain whitefish *Prosopium williamsoni* (MW), northern pikeminnow (formerly northern squawfish) *Ptychocheilus oregonensis* (NS), peamouth *Mylocheilus caurinus* (PC), prickly sculpin *Cottus asper* (PS), rainbow trout *O. mykiss* (RT), and redside shiner *Richardsonius balteatus* (RS). Vertical distance from each species to 1:1 line represents the expected model performance bias resulting from using a resubstitution approach to model validation.
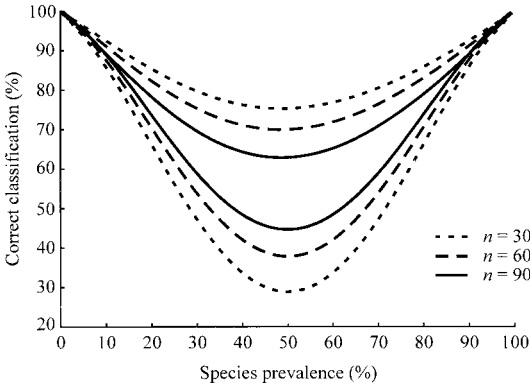
FIGURE 3.—Results from the Monte Carlo experiment showing the relationship between species prevalence (i.e., frequency of species occurrence) and chance correct classification based on data exhibiting random associations between the dependent and independent variables. Results are based on 10,000 simulated data sets and illustrate the 95% confidence intervals for 3 sample sizes (n = 30, 60, and 90).

whelming importance of testing model predictions against chance predictions produced solely by the artifact of species prevalence in the data.

Empirical Data

Given the inherent dependence of the correct classification rate on species prevalence, it is im-

portant to either quantify model performance with a measure that is invariant to the effects of species prevalence, or test the significance of the classification rate against expectations based on chance. In the first case, there exist a number of alternative measures that quantify model performance, including the odds ratio, the normalized mutual information statistic, Huberty's normal statistic, Cohen's kappa, and the area under the curve from receiver operating characteristics plots (see Fielding and Bell 1997 for a review). Recently, Manel et al. (2001) provided a comprehensive comparison of a number of these measures using an extensive empirical data set, finding Cohen's kappa to be the most robust measure as it was only marginally affected by prevalence. We have employed the second case approach by using a randomization method to assess whether the correct classification rates of the Blackwater River and Petawawa River fish habitat models differed significantly from chance (e.g., Solow 1990). A null distribution of correct classification rates (CCRs) for each species was generated by randomly permuting the original observations of occurrences among the lake or stream sites, conducting logistic regression analysis using the randomized species occurrence and the original independent variables, and calculating
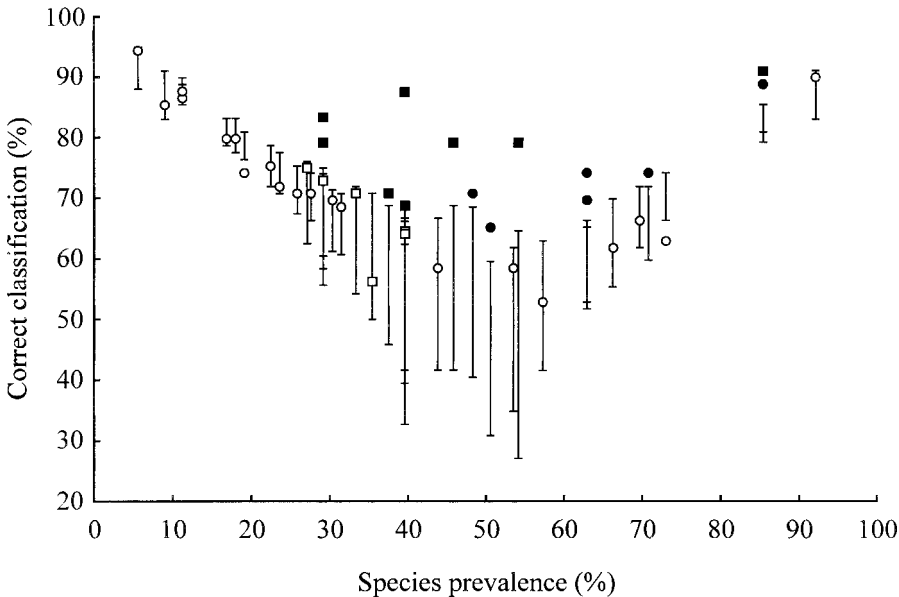

FIGURE 4.—Results from the randomization test for assessing model predictive performance. Whisker plots show the 95% confidence intervals based on 999 randomizations of the data set. Observed correct classification rates are represented by circles (species from the Petawawa River drainage) and squares (species from the Blackwater River drainage), where solid symbols represent fish habitat models whose predictions statistically differ from chance ($P < 0.05$) and open symbols represent nonsignificant predictions.

the jackknifed CCR. This procedure was repeated 999 times and the significance level of the predictive model was calculated as the proportion of random CCRs (including the observed CCR) that were larger than or equal to the observed CCR. Figure 4 shows the 95% confidence intervals of the correct classification rates based on the randomized correct classification rates and the observed correct classification rate for the Petawawa River and Blackwater River species, which are plotted as a function of their prevalence in the data sets. Observed correct classification rates located above the 95% confidence interval represent fish habitat models whose predictions differ statistically from random. The results from the randomization test show that for over half the species modeled (18 out of 26 Petawawa species and 5 out of the 14 Blackwater species), the classification success for predicting species presence or absence was not significantly different from random (Figure 4). For example, the two least prevalent species, blackchin shiner and rock bass, were predicted correctly in 94% and 85% of the lakes in the Petawawa River drainage, respectively; yet these values were no better than expected due to chance (i.e., 94% and 91%, respectively). Similarly, the most prevalent species, white sucker, had an impressive correct classification of 90%; however, this model failed to produce predictions greater than the 92% expected by chance. Close agreement can also be noted between the chance 95% confidence intervals of correct classifications and species prevalence from Figure 4 and the pattern shown in Figure 3.

Consequences of Chance Predictions

Our results show that a model may produce very high correct classification rates in extreme situations, for instance where species prevalence is very low (<10%) or very high (>90%), but its improvement over chance may be slight. In such cases, the random classification of species as being present or absent based on its a priori probability of occurrence (i.e., species prevalence) provides comparable rates of correct classification relative to the predictions generated by the logistic models. The effect of species prevalence on the estimated classification success of a model is unavoidable because, with increased frequency of occurrence, there is a greater probability of correctly predicting the species to be present. As a result, one should expect increasing model sensitivity (the ability to correctly predict species presence) and decreasing model specificity (the ability to correctly predict

species absence) to occur with increasing species prevalence, and that overall classification success should be highest for rare and common species and decrease as species prevalence approaches 50% (see Manel et al. 1999; Olden and Jackson 2001). Although high correct classification rates might at first appear attractive and suggest useful models, we show that chance predictions can lead to false confidence in these models for predicting species distributions. For this reason, it is clearly necessary that the classification rates of predictive models be tested against expectations based on chance prior to the interpretation and future application of such models.

Conclusion

Our primary goal was to enhance the awareness of fisheries biologists to the dangers of commonly employed model validation techniques and the phenomena of chance predictions in statistical modeling. Researchers should be aware of the magnitude of the biases that are associated with resubstitution estimates of model predictive performance as they can result in an overly optimistic representation of the model's ability and can lead to potentially unwarranted and misleading conclusions. It is important to note that although we used only logistic regression in this study, our findings are relevant for other statistical approaches such as linear or multiple linear regression, discriminant analysis, classification trees, and artificial neural networks, which, in some cases, may exhibit even greater model performance biases (J. Olden, unpublished data). After model predictions are produced, it is necessary that they be tested against predictions based on chance alone in addition to subjecting them to proper validation. In this study, we demonstrated the effect of species prevalence on rates of chance predictions, and the influence of this relationship on our confidence, or lack thereof, in model predictions.

In conclusion, although predictive models serve as valuable tools in the fisheries sciences, researchers must increase their emphasis on the proper validation and assessment of predictions generated from these models. The failure to properly address these issues will only weaken the potential utility of such models and ultimately will result in models of limited usefulness. To enable the confident application of predictive models in the management and conservation of our fisheries resources, we appeal to researchers to employ proper and defensible methodologies for model validation and assessment.

References

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A 158:419–466.

Crossman, E. J., and N. E. Mandrak. 1992. 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.

Efron, B. 1986. How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81:461–470.

Evans, D. O., and C. H. Olver. 1995. Introduction of lake trout (*Salvelinus namaycush*) to inland lakes of Ontario, Canada: factors contributing to successful colonization. Journal of Great Lakes Research 21(Supplement 1):30–53.

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.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

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

Hawkins, C. P., R. H. Norris, J. N. Hogue, and J. W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10: 1456–1477.

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., 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.

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.

Manel, S., H. C. Willia ms, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38:921–931.

Oberdorff, T., D. Pont, B. Hugueny, and D. Chessel. 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. Freshwater Biology 46: 399–415.

Olden, J. D. 2000. Predictive models for freshwater fish community composition. Master's Thesis. University of Toronto, Toronto.

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., and D. A. Jackson. 2001. Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. Transactions of the American Fisheries Society 130: 878–897.

Peterson, A. T., and D. A. Vieglais. 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attacking a pressing problem. BioScience 51:363–371.

Porter, M. S., J. Rosenfeld, and E. A. Parkinson. 2000. Predictive models of fish species distribution in the Blackwater drainage, British Columbia. North American Journal of Fisheries Management 20: 349–359.

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

Solow, A. R. 1990. A randomization test for misclassification probability in discriminant analysis. Ecology 71:2379–2382.

Stefan, H. G., X. Fang, and J. G. Eaton. 2001. Simulated fish habitat changes in North American lakes in response to projected climate change. Transactions of the American Fisheries Society 130:459–477.

Verbyla, D. L., and J. A. Litaitis. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. Environmental Management 13:783–787.

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: 1989. Fisheries 14(6):2–20.

Williams, P. H., and M. B. Araujo. 2000. Using probability of persistence to identify important areas for biodiversity conservation. Proceedings of the Royal Society of London B 267:1959–1966.