

available at [www.sciencedirect.com](http://www.sciencedirect.com)[www.elsevier.com/locate/ecolinf](http://www.elsevier.com/locate/ecolinf)

## Informatics software for the ecologist's toolbox: A basic example

J.B. Williams\*, N.L. Poff

Department of Biology, Colorado State University, Fort Collins, CO 80523, USA

### ARTICLE INFO

#### Article history:

Received 19 February 2005

Received in revised form

7 March 2006

Accepted 15 March 2006

#### Keywords:

Artificial neural networks

Genetic algorithms

Classification trees

Software

Fish IBI

### ABSTRACT

Machine learning techniques for ecological applications or “eco-informatics” are becoming increasingly useful and accessible as software for these techniques becomes more readily available. Complex ecological data sets with a multitude of variables are also increasingly available. Ecologists, who do not necessarily have extensive backgrounds in machine-learning techniques, are facing decisions on new methods of data analysis. We evaluated the predictive ability of three commercially available (i.e. user-friendly) software packages for artificial neural networks (ANNs), evolutionary algorithms (EAs), and classification/regression trees (CART). To demonstrate their usage, we analyzed fish and habitat data from the mid-Atlantic region of the US, which was collected by the U.S. Environmental Protection Agency (EPA). These data, including over 200 environmental descriptors summarizing watershed, stream, and water chemistry, and physical habitat characteristics in addition to fish community metrics (i.e. richness, Index of Biotic Integrity (IBI) scores, % exotics), were collected as part of the EPA's Environmental Monitoring and Protection program. We predicted fish IBI scores as a function of these local and regional scale habitat variables. Predictive ability is evaluated with independent validation data. These approaches could prove especially useful for conservation or management applications where ecologists seek to utilize the most comprehensive data to make predictions at various scales. By employing “user-friendly” software we hope to show that ecologists, without extensive knowledge of computational science, can benefit from these techniques by extracting more information about complex ecosystems. We found that all models predicted better than chance ( $p < 0.05$ ). Relative strengths and weaknesses of these three approaches are compared and recommendations for their use in ecological applications are presented.

© 2006 Elsevier B.V. All rights reserved.

### 1. Introduction

Ecological modeling is important in order to understand and describe natural phenomena by economizing the thought required to interpret ecological data. An effective model links the data to ecological questions and provides a sufficient amount of understanding or predictions, where

an ecologist's perception alone falls short. Advances in data collection technology, such as GPS, remote-sensing, and larger/faster computers, have led to the increased availability of large, complex ecological data sets. As technology for data collection has advanced, so has analytical technology. Informatics approaches, such as Artificial Neural Networks, Evolutionary Algorithms, and Decision Trees, have shown

\* Corresponding author.

E-mail addresses: [jwilli@lamar.colostate.edu](mailto:jwilli@lamar.colostate.edu) (J.B. Williams), [poff@lamar.colostate.edu](mailto:poff@lamar.colostate.edu) (N.L. Poff).

excellent ability for the advancement of ecological modeling (De'ath and Fabricius, 2000). These “eco-informatics” approaches have several advantages over traditional ecological models (such as various regression techniques), which are often limited because of assumptions of normality, obligatory transformations, and ineptness with few or more zero values (Lek et al., 1996). Eco-informatics models are not limited by these drawbacks and have consistently out-performed traditional approaches when analyzing equivalent non-linear data sets (Lek et al., 1996; Olden and Jackson, 2001, 2002a).

New advances in computation sciences have led to the availability of “informatics tools” for ecological applications. Many ecological researchers have begun to utilize the abilities of informatics approaches for widespread applications such as predicting patterns of species richness (Guegan et al., 1998), multivariate analysis of biological invasions (Kolar and Lodge, 2002), and predicting phytoplankton abundance in freshwater lakes from time series data (Wigham and Recknagel, 2001). However, eco-informatics approaches have not yet been widely embraced into the “ecologist’s toolbox”. One reason for this lag is that some ecologists lack computational background needed to operate certain software implementations (Fielding, 1999). Many ecologists do not have a background in computational science and may be hesitant to invest their time in learning extensive program code language and syntax. However, as the popularity of these approaches grows, there is new software available. We propose that the eco-informatics community must encourage the use and standardization of this “user-friendly” software for ecological applications. Such software will increase informatics usage and awareness among ecologists, and promote the advance of these analytical methods.

We will demonstrate and evaluate the ability of these approaches by predicting fish biotic integrity in the Mid-Atlantic Highlands Area (MAHA) of the eastern United States. We use a stratified sample data set available for the region collected by the US Environmental Protection Agency (EPA) which contains over 200 environmental variables including watershed, and water chemistry descriptors (Herlihy et al., 2000). The EPA also collected physical habitat and riparian disturbance at a smaller subset of sites (McCormick et al., 2001). This allows us the opportunity to evaluate the ability of informatics methods with the challenging task, using a relatively small data set to predict stream biotic integrity with a sizeable set of predictor variables. We evaluated the predictive ability of three commercially available (i.e. user-friendly) software packages for artificial neural networks (ANNs), evolutionary algorithms (EAs), and classification/regression trees (CART). We refer the reader to Fielding (1999) for a detailed introduction to each of these methods.

### 1.1. MAHA

In 1993 the US Environmental Protection Agency (EPA) developed a randomized sample design for the Mid-Atlantic Highlands Area (MAHA), which extend from the Virginia/North Carolina border to the Catskill Mountains of New York (Herlihy et al., 1993), as part of the Environmental Monitoring and Assessment Program (EMAP; Herlihy et al., 2000). Fish

assemblage and environmental data was collected at 309 sites, which represent Wadeable streams within the MAHA study region (Lazorchak et al., 1998). A smaller subset of sites were also measured for physical habitat and riparian disturbance (Kaufmann and Robison, 1998). These local scale physical factors have been shown to be important in determining biotic integrity (Snyder et al., 2003). The Mid-Atlantic Highlands area has a variety of stream conditions due to several environmental stressors. Significant portions of the MAHA area have poor or impaired conditions with respect to total nitrogen (5% of total stream miles), total phosphorus (5%), fish tissue contamination (10%), acidic deposition (11%), mine drainage (14%), riparian disturbance (24%), and channel sedimentation (25%), based on EPA criteria (EPA, 2000). This wide range of environmental stressors makes the MAHA area a good candidate for analysis of environmental/biotic relationships (Herlihy et al., 1998).

### 1.2. IBI

The Index of Biotic Integrity (IBI) is a composite index that uses combined metrics for biological attributes (ecological, trophic, and reproductive) of fish communities that best explain water quality relative to reference sites for specific regions (Hughes et al., 1998). The status of the fish community, as represented by an IBI score, can be used as a biological indicator of environmental stressors (Fausch et al., 1984). McCormick et al. (2001) developed an IBI specifically for the MAHA area and

**Table 1 – Environmental variables selected for use in predictive IBI models including watershed scale, reach scale, and water chemistry variables**

Reach scale	
ELEV	Elevation of stream index site (m)
ORDER	Stream order
PCT_SAFN	Substrate sand and fines <2 mm (%)
SINU	Channel sinuosity (m/m)
W1_HALL	Riparian disturbance—sum all types
WD_RAT	Mean width/depth ratio
XCDENBK	Mean bankside canopy density (%)
XEMBED	Mean embeddedness—channel and margin (%)
XFC_NAT	Fish cover—natural types (sum area proportion)
XSLOPE	Channel slope—reach mean (%)
Watershed scale	
AREA_WS	Watershed area digitized from maps
FOR_TOT	% watershed—forest
RD_DEN	Road density (m/ha)
URB_TOT	% watershed—urban lands
AG_TOT	% watershed—agricultural lands
MIN E_TOT	% watershed—strip mines/quarries/gravel pits
Water chemistry	
ANC	Acid neutralizing capacity (µeq/L)
CL	Chloride (µeq/L)
NH4	Ammonium (µeq/L)
NTL	Total nitrogen (µg/L)
PTL	Total phosphorus (µg/L)
SO4	Sulfate (µeq/L)

The left column is the variable code as designated in the EMAP data set, the right column is a brief explanation of the variable (Lazorchak et al., 1998).

**Table 2 – Results for predictive IBI models with artificial neural networks (ANNs), evolutionary algorithms (EAs), and tree-based (CART) models**

	Validation			Training		
	CCR	Kappa	p	CCR	Kappa	p
ANN	0.4125	0.40032	<0.0001	0.73438	0.60427	<0.0001
CART	0.375	0.12146	0.044	0.725	0.60136	<0.0001
EA excellent	0.9	0.65517	<0.0001	0.875	0.59103	<0.0001
EA good	0.8	0.38462	<0.01	0.85938	0.52312	<0.0001
EA fair	0.8125	0.49664	<0.001	0.79063	0.431	<0.0001
EA poor	0.8625	0.61404	<0.0001	0.875	0.66015	<0.0001

For EA models, four independent models were used to predict each IBI class (excellent, good, fair, and poor) vs. the alternative. Model accuracy is evaluated with the correct classifications rate (CCR) and Kappa statistic, which can be interpreted as the percent improvement over chance-alone predictions. The p-value is listed for a test of significance (Z), with  $H_0: K=0$ . All statistics are reported separately for validation and training data sets.

assigned the IBI scores to 4 classes. They found that 5% of MAHA streams had an IBI of excellent, 22% were good, and 38% and 14% of streams were fair or poor, respectively; 21% of streams were not assessed for IBI.

## 2. Methods

Environmental data were compiled into a database using unique identification codes. For any sites with multiple visits, only the first visit was kept in the database. We also eliminated sites that were not sampled for physical habitat variables. Eight sites were removed that had one or more missing variable values, leaving a total of 80 sites. Because we were limited to 80 sites, we chose to use an  $n-1$  cross-validation (Fielding, 1999). The sites were randomly divided into five groups, and each experiment was performed five times using each one of the five groups as validation data and the remaining 4 groups as training data. Combined results from five models were used to determine overall model correct classification rate (CCR). We also calculated Cohen’s Kappa statistic, the chance corrected percentage of

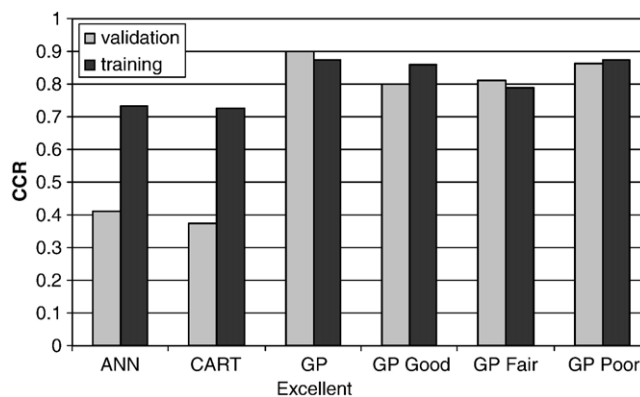
agreement between predicted and observed classes (Titus et al., 1984).

Over 200 quantitative variables were available that summarized water chemistry, land use, and riparian habitat. The initial list of candidate water chemistry variables were selected from a list of water quality indicators shown to have a correlation ( $|r|>0.15$ ) with nine individual MAHA IBI metrics (see McCormick et al., 2001). We also selected candidate reach and watershed scale variables that quantify a variety of anthropogenic disturbances and other available variables deemed important to fish communities from literature review and expert judgment. Next, all variables were analyzed using a correlation table, and for any values with  $|r|>0.5$ , we removed one of the correlated variables. Finally, the number variables were further reduced to 22 based on importance to fish (from literature review and expert judgment), and interpretability (Table 1).

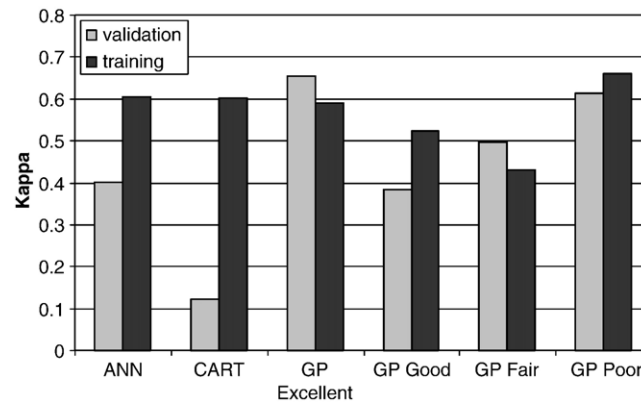
Software for model creation was selected for ease of use and interpretability. We used EasyNN (version 4.0b) by Neural Planner Software, Discipulus (demo version 3) by RML technologies, and CART (version 5.0) by Salford Systems for artificial neural networks (ANNs), evolutionary algorithms (EAs), and classification/regression trees (CART) models, respectively. All models were created using software defaults unless otherwise specified. For EasyNN, we set the number of hidden layers to one. For Discipulus we decomposed the classification into 4 separate models because the software is limited to classification problems with only two classes. The four models independently predicted each IBI class (excellent, good, fair, and poor) vs. the alternative. CART was set to create the best tree by minimizing classification error.

## 3. Results

The results summarized in Table 2 show that all models predicted better than chance ( $p<0.05$ ), with kappa scores ranging from 0.121 to 0.660. The correct classification of the training data set was above 70% for all models. However, the ANN and CART models showed inferior classification performance with the validation data set (Fig. 1). The ANN model performed 40% better than chance (kappa=0.400,  $p<0.0001$ ),



**Fig. 1 – Correct classification rate (CCR) for predictive IBI models using artificial neural networks (ANNs), evolutionary algorithms (EAs), and tree-based (CART) models. For EA models, four independent models were used to predict each IBI class (excellent, good, fair, and poor) vs. the alternative.**



**Fig. 2**–Cohen's Kappa statistic for predictive IBI models using artificial neural networks (ANNs), evolutionary algorithms (EAs), and tree-based (CART) models. For EA models, four independent models were used to predict each IBI class (excellent, good, fair, and poor) vs. the alternative. For predictive IBI models artificial neural networks (ANNs), evolutionary algorithms (EAs), and tree-based (CART) models were used. For EA models, four independent models were used to predict each IBI class (excellent, good, fair, and poor) vs. the alternative. The Kappa statistic can be interpreted as the percent improvement over chance-alone predictions.

but the CART model had a low correct classification rate, which was only slightly significant when compared to chance predictions (CCR=0.375, kappa=0.121,  $p=0.044$ ). All EA models achieved at least 79% correct classification; the training and validation data sets performed similarly. In both data sets, the extreme IBI class models (excellent and poor) performed better than the intermediate IBI class models (Fig. 2).

#### 4. Discussion

Several advantages of CART for ecological applications have been recognized including: flexibility for numeric or categorical dependant variables, capacity to explore data interactions, and easy graphical interpretation (De'ath and Fabricius, 2000). The application of neural networks to problems in ecology has increased lately, in large part because of their perceived ability to accurately model complex data (Lek and Guégan, 1999). ANN allows variables to interact positively or negatively and at a variety of strengths simultaneously. It is this flexibility that has likely led to the increased popularity of neural networks in ecology. EAs have advantages over traditional ecological modeling because they are robust to non-linear data, uneven sampling, and small sample sizes (Peterson and Cohoon, 1999). Genetic Programming has had "startling empirical success" in other fields (Whitley, 2001) and there is still great potential in ecology. Overall, EAs are very powerful and have been used in a number of ecological applications, yet they are often difficult to interpret (Wigham and Recknagel, 2001). Although not demonstrated here, CART and ANN both have the advantage of allowing for interpretable graphic visualizations of models (De'ath and Fabricius, 2000; Olden and Jackson, 2002b).

Although this is a limited study, it shows that informatics tools can be applied using available software. This "user-friendly" software will appeal to a larger audience of ecologists. Though not presented here, these approaches also hold potential for use in exploratory analysis of data and variable importance in predictions. However, more research is needed

to develop a protocol for employing such techniques. Many seemingly arbitrary choices for model parameters and use of software default settings may have important ramifications for experimental results. We also found that certain software may have limitations that make results difficult to interpret. For example, the EA model seems to predict quite well, but the results may be misleading. Because it is easier to classify data into 2 lumped classes than 4 classes, the results are not directly comparable to the other models. However, if predicting a certain class (poor IBI for example) is of primary concern, this method of decomposing the classification into 2 classes may be useful. CART and ANN performed approximately equally well, in both the training and validation data sets. Given the small number of sites and large number of predictor variables, the models performed reasonably well. The accuracy might be improved with a more detailed analysis. Overall, we demonstrated that these tools can be utilized straightforwardly to make ecological predictions.

#### REFERENCES

- De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81, 3178–3192.
- EPA, 2000. Mid-Atlantic Streams Assessment. EPA/903/R-00/015. US Environmental Protection Agency Region 3. Philadelphia, PA.
- Fausch, K.D., Karr, J.R., Yant, P.R., 1984. Regional application of an index of biotic integrity based on stream fish communities. *Transactions of the American Fisheries Society* 113, 39–55.
- Fielding, S.H. (Ed.), 1999. *Machine Learning Methods for Ecological Applications*. Kluwer Academic, Dordrecht, The Netherlands.
- Guegan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. *Nature* 391, 382–384.
- Herlihy, A.T., Kaufmann, P.R., Church, M.R., Wightingon, P.J., Webb, J.R., Sale, M.J., 1993. The effects of acidic deposition on

- streams in the Appalachian Mountain and Piedmont Region of the Mid-Atlantic United States. *Water Resources Research* 29, 2687–2703.
- Herlihy, A.T., Stoddard, J.L., Johnson, C.B., 1998. The relationship between stream chemistry and watershed land cover data in the mid-Atlantic region, US. *Water, Air and Soil Pollution* 105, 377–386.
- Herlihy, A.T., Larsen, D.P., Paulsen, S.G., Urquhart, N.S., Rosenbaum, B.J., 2000. Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP Mid-Atlantic Pilot Study. *Environmental Monitoring and Assessment* 63, 95–113.
- Hughes, R.M., Kaufmann, P.R., Herlihy, A.T., Kincaid, T.M., Reynolds, L., Larsen, D.P., 1998. A process for developing and evaluating indices of fish assemblage integrity. *Canadian Journal of Fisheries and Aquatic Sciences* 55, 1618–1631.
- Kaufmann, P.R., Robison, G., 1998. Physical habitat characterization. In: Lazorchak, J.M., Klemm, D.L., Peck, D.V. (Eds.), *Environmental Monitoring and Assessment Program Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams*. U.S. Environmental Protection Agency, Washington, DC, pp. 77–118. EPA/620/R-94/004F.
- Kolar, C.S., Lodge, D.M., 2002. Ecological predictions and risk assessment for alien fishes in North America. *Science* 298, 1233–1236.
- Lazorchak, J.M., Klemm, D.L., Peck, D.V., 1998. *Environmental Monitoring and Assessment Program Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams*. U.S. Environmental Protection Agency, Washington, DC. EPA/620/R-94/004F.
- Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* 120, 65–73.
- Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* 90, 39–52.
- McCormick, F.H., Hughes, R.M., Kaufmann, P.R., Peck, D.V., Stoddard, J.L., Herlihy, A.T., 2001. Development of an index of biotic integrity for the Mid-Atlantic Highlands region. *Transactions of the American Fisheries Society* 130, 857–877.
- Olden, J.D., Jackson, D.A., 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.
- Olden, J.D., Jackson, D.A., 2002a. A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology* 47, 1976–1995.
- Olden, J.D., Jackson, D.A., 2002b. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154, 135–150.
- Peterson, A.T., Cohoon, K.P., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling* 117, 159–164.
- Snyder, C.D., Young, J.A., Vilella, R., Lemarie, D.P., 2003. Influences of upland and riparian land use patterns on stream biotic integrity. *Landscape Ecology* 18, 647–664.
- Titus, K., Mosher, J.A., Williams, B.K., 1984. Chance-corrected classification for use in discriminant analysis—ecological applications. *American Midland Naturalist* 111, 1–7.
- Whitley, D., 2001. An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and Software Technology* 43, 817–831.
- Wigham, P.A., Recknagel, F., 2001. An inductive approach to ecological time series modelling by evolutionary computation. *Ecological Modelling* 146, 275–287.