Fisheries Research 107 (2011) 261–271

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Decreasing uncertainty in catch rate analyses using Delta-AdaBoost: An alternative approach in catch and bycatch analyses with high percentage of zeros

Yan Li*, Yan Jiao, Qing He

Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0321, USA

ARTICLE INFO

Article history:
Received 5 July 2010
Received in revised form 27 October 2010
Accepted 9 November 2010

Keywords:
Delta model
AdaBoost
Tweedie distribution
Catch rate
Zero catch

ABSTRACT

The gillnet data of walleye (*Sander vitreus*), yellow perch (*Perca flavescens*), and white perch (*Morone americana*), collected by a fishery-independent survey (Lake Eire Partnership Index Fishing Survey, PIS) from 1989 to 2008, contained 75–83% of zero observations. AdaBoost algorithm was applied to the model analyses with such fishery data for each species. The 3- and 5-fold cross-validations were conducted to evaluate the performance of each candidate model. The performance of the delta model consisting of one generalized additive model and one AdaBoost model (Delta-AdaBoost) was compared with five candidate models. The five candidate models included: the delta model comprising two generalized linear models (Delta-GLM), the delta model comprising two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly), the delta model comprising two generalized additive models (Delta-GAM), the generalized linear model with Tweedie distribution (GLM-Tweedie), and the generalized additive model with Tweedie distribution (GAM-Tweedie). To predict the presence/absence of fish species, the performance of AdaBoost model was compared in terms of error rate with conventional generalized linear and additive models assuming a binomial distribution. Results from 3- and 5-fold cross-validation indicated that Delta-AdaBoost model yielded the smallest training error (0.431–0.433 for walleye, 0.528–0.519 for yellow perch and 0.251 for white perch) and test error (0.435–0.436 for walleye, 0.524 for yellow perch and 0.254–0.255 for white perch) on average, followed by Delta-GLM-Poly model for yellow perch and white perch, and Delta-GAM model for walleye. In the prediction of the presence/absence of fish species, AdaBoost model had the lowest error rate, compared with generalized linear and additive models. We suggested AdaBoost algorithm to be an alternative to deal with the high percentage of zero observations in the catch and bycatch analyses in fisheries studies.

1. Introduction

Catch and bycatch rate estimations play an indispensable role in fish stock assessment and management (Gunderson, 1993; Helser and Hayes, 1995; Maunder and Punt, 2004). Various methods have been developed to estimate the catch and bycatch rates for a specific fishery. The commonly used methods include the ratio method, which determines catch rates relative to a standard value (Beverton and Holt, 1957); the generalized linear model, which incorporates multiple variables to describe the environmental and fishing effects (Gavaris, 1980; Kimura, 1981); and the generalized additive model, which demonstrates the nonlinear relationship between the catch/bycatch rate and explanatory variables through a smooth function (Bigelow et al., 1999; Damalas et al., 2007). However, these methods have difficulties in dealing with the highly skewed data where a large amount of zeros are included. Such data are frequently encountered in the catch analyses of rare species and the bycatch analyses (Maunder and Punt, 2004; Ortiz et al., 2000). The presence of zeros may invalidate the assumptions of normality we usually use, and may cause computational difficulties.

Ignorance of a considerable proportion of zeros may result in a loss of information that reflects the spatial or temporal distribution characteristics of fish stocks. Two types of approaches have been applied in previous studies to deal with zeros in fishery data analyses. One approach is to add a small constant to each zero observation of the response variable, followed by a generalized linear or additive model analysis (Maunder and Punt, 2004; Ortiz et al., 2000; Shono, 2008). However, the estimation results are sensitive to the choice of the constant (Maunder and Punt, 2004; Ortiz et al., 2000). The other approach is to utilize the delta model and the Tweedie distribution model. In the delta model, the positive values are fitted by a generalized linear or additive model, and the probabilities of observing zero values are fitted by a generalized linear or additive model with an assumption of binomial

* Corresponding author at: Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, 100 Cheatham Hall, Blacksburg, VA 24061-0321, USA. Tel.: +1 540 8088373.
E-mail address: yanli08@vt.edu (Y. Li).

0165-7836/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2010.11.008

distribution (Lo et al., 1992; Maunder and Langley, 2004; Stefansson, 1996; Ye et al., 2001). Although combining two sub-models complicates the model interpretation in that the explanatory variables may differ in two sub-models, the delta model has been widely used to estimate bycatch (Murray, 2004; Ortiz et al., 2000), catch rate, and abundance index (Lo et al., 1992; Stefansson, 1996; Ye et al., 2001). By contrast, the Tweedie distribution model handles zero data uniformly along with the positive data, where the Tweedie distribution is considered to be a Poisson–Gamma compound distribution when its power parameter is greater than 1 but less than 2 (Shono, 2008; Tweedie, 1984). The Tweedie distribution model has been judged to outperform the generalized linear model with an additive constant and the delta model composed of two generalized linear models (Shono, 2008).

AdaBoost is a typical boosting algorithm that was originally used for classification problems. The algorithm used for classification is called a classifier. The final strong classifier is obtained by successively applying a classification algorithm to reweighted data and then combining a sequence of weak classifiers that minimize the prediction error at each iteration (Freund and Schapire, 1996; Friedman et al., 2000; Hastie et al., 2001; Kawakita et al., 2005). In a fishery context, zeros and positive values can be converted into a categorical variable $\{-1,1\}$, indicating the events of no fish caught and the events of at least one fish caught, respectively, and then the problem can be treated as a two-group classification problem (Kawakita et al., 2005). This method has been used to predict the occurrence of large silky shark bycatch in a tuna purse-seine fishery, and the results confirmed the superiority of AdaBoost model in bycatch analyses where data were skewed by zeros (Kawakita et al., 2005).

The present study was based on the gillnet data collected from a fishery-independent survey, the Lake Erie Partnership Index Fishing Survey (PIS). The PIS survey was primarily operated by the Ontario Ministry of Natural Resources (OMNR) and the Ontario Commercial Fisheries Association (OCFA) since 1989. The experimental gillnets with mesh size ranging from 32 to 152 mm were deployed across the Ontario waters of Lake Erie in the fall (August–November) annually, using commercial fishing vessels and commercial fishing crews (OCFA, 2007).

We focused on three species in this analysis, walleye (*Sander vitreus*), yellow perch (*Perca flavescens*), and white perch (*Morone americana*). Walleye and yellow perch dominate the commercial gillnet fisheries in Lake Erie (Kinnunen, 2003; Thomas and Haas, 2005), and white perch has imposed considerable influences on fish communities and the lake ecosystem as an invasive species (Parrish and Margraf, 1990; Schaeffer and Margraf, 1987; Scott and Crossman, 1973).

In the present study, the delta model comprising one generalized additive model and one AdaBoost model (Delta-AdaBoost) was developed to estimate the catch rates of walleye, yellow perch and white perch based on the PIS data from 1989 to 2008. The performance of the Delta-AdaBoost model was compared with five candidate models, including the delta model comprising two generalized linear models (Delta-GLM), the delta model comprising two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly), the delta model comprising two generalized additive models (Delta-GAM), the generalized linear model with Tweedie distribution (GLM-Tweedie), and the generalized additive model with Tweedie distribution (GAM-Tweedie). The performance of the AdaBoost model to predict the presence/absence of fish species was compared in terms of error rate with the generalized linear and additive model assuming a binomial distribution. Each model was evaluated through the 3- and 5-fold cross-validation. The goals of this study were (1) to evaluate the performance of the Delta-AdaBoost model and the AdaBoost model in analyzing the fishery data with high percentage of zeros and (2) to explore the application of AdaBoost algorithm in fishery studies.

2. Methods

2.1. Data and variables

We estimated the catch rates of walleye, yellow perch, and white perch using the PIS data from 1989 to 2008 provided by OCFA. The PIS data included a large number of zero observations (75–83%), and as a result, the commonly used assumption on normal or lognormal distribution was violated (Ortiz et al., 2000). Totally 53,662 records were available for analysis and the catch rate was expressed as catch in weight (kg) per net (30.5 m long × 1.8 m deep).

Fourteen explanatory variables were available: nine continuous variables, i.e., site depth, gear depth, secchi depth, gear temperature, dissolved oxygen, soak time, site temperature, longitude, and latitude; five categorical variables, i.e., basin (five basins), year (twenty years), month (four months), gear type (two gear types), and mesh size (fourteen mesh sizes). Site temperature is the water surface temperature. Gear temperature means the water temperature at the gear set depth. Gear type refers to canned or bottomed gillnets.

The correlation coefficients among all explanatory variables were examined to detect those that were highly correlated. A preliminary stepwise selection based on Akaike Information Criterion (AIC, Akaike, 1974) was further conducted to eliminate one of the correlated pair of variables, i.e., the variable that yielded a larger AIC value was eliminated from the correlated pair. The remaining variables were selected through a stepwise procedure based on AIC (Akaike, 1974; Burnham and Anderson, 2002). The model with smaller AIC was considered to fit the data better. Interaction terms were not included in the regression model to avoid additional multicollinearity problems and difficulties in model interpretation (Damalas et al., 2007; Maunder and Punt, 2004).

2.2. Delta model and Delta-AdaBoost model

A delta model usually consists of two components, one model to fit the positive values and the other to estimate the probability of obtaining non-zero captures. Estimates of the catch rate from a delta model can be obtained by multiplying these two components (Lo et al., 1992; Maunder and Punt, 2004; Murray, 2004; Ortiz et al., 2000; Pennington, 1996; Stefansson, 1996; Ye et al., 2001):

$$Catch\ \ \hat{rate} = \hat{d} \times \hat{q},$$

where $Catch\ \ \hat{rate}$ is the estimate of catch rate, $\hat{d}$ is the estimate of catch rate when only positive values of the response variable are analyzed, and $\hat{q}$ is the estimate of the probability of obtaining non-zero captures.

In the delta model, the model to fit the positive values could be a generalized linear model (Eq. (1)), a generalized linear model with polynomial terms up to degree 3 (Eq. (2)), or a generalized additive model (Eq. (3)), which were built by assuming a lognormal distribution as follows:

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} \beta_j X_j, \tag{1}$$

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} (\beta'_j X_j + \beta''_j X_j^2 + \beta'''_j X_j^3), \tag{2}$$

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} f_j(X_j), \tag{3}$$