ORIGINAL ARTICLE

http://dx.doi.org/10.1590/S2675-28242020068226

OCEAN AND COASTAL RESEARCH

Standardizing catch per unit effort by machine learning techniques in longline fisheries: a case study of bigeye tuna in the Atlantic Ocean

Shenglong Yang[1,2], Yang Dai[2], Wei Fan[2]*, Huiming Shi[1]

[1] Key Laboratory of Oceanic and Polar Fisheries, Ministry of Agriculture and Rural Affairs; East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shanghai, 200090, China
[2] Key and Open Laboratory of Remote Sensing Information Technology in Fishing Resource, East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shanghai 200090, China

*Corresponding author: fanwee@126.com

ABSTRACT

Support vector machine (SVM) is shown to have better performance in catch per unit of effort (CPUE) standardization than other methods. The SVM performance highly relates to its parameters selection and has not been discussed in CPUE standardization. Analyzing the influence of parameter selection on SVM performance for CPUE standardization could improve model construction and performance, and thus provide useful information to stock assessment and management. We applied SVM to standardize longline catch per unit fishing effort of fishery data for bigeye tuna (*Thunnus obesus*) in the tropical fishing area of Atlantic Ocean and evaluated three parameters optimization methods: a Grid Search method, and two improved hybrid algorithms, namely SVMs in combination with the particle swarm optimization (PSO-SVM), and genetic algorithms (GA-SVM), in order to increase the strength of SVM. The mean absolute error (MAE), mean square error (MSE), three types of correlation coefficients and the normalized mean square error (NMSE) were computed to compare the algorithm performances. The PSO-SVM and GA-SVM algorithms had particularly high performances of indicative values in the training data and dataset, and the performances of PSO-SVM were marginally better than GA-SVM. The Grid search algorithm had best performances of indicative values in testing data. In general, PSO was appropriate to optimize the SVM parameters in CPUE standardization. The standardized CPUE was unstable and low from 2007 to 2011, increased during 2011-2013, then decreased from 2015 to 2017. The abundance index was lower compared with before 2000 and showed a decreasing trend in recent years.

Descriptors: CPUE standardization; Support Vector Machine; Grid Search; Genetic Algorithms; Particle Swarm Optimization.

INTRODUCTION

Fisheries play a significant role in the global food supply (Garcia and Rosenberg, 2010). The catch per unit of effort (CPUE) is used as a relative abundance of fishery resources (Ricker, 1975; Bigelow et al., 1999) and plays an important role in resource assessment and management (Maunder and Langley, 2004; Pauly et al., 2013). CPUE is estimated mainly from commercial or recreational fisheries and requires time-consuming and costly data collection (Maunder and Punt, 2004; Ward et al., 2013). The nominal CPUEs derived from such data are greatly influenced by spatial, temporal and environmental factors, among others. Therefore, CPUE derived directly from raw fishing data needs to be standardized using statistical models to remove those effects (Maunder and Punt, 2004).

Several methods have been applied in CPUE standardization (Guan et al., 2014), including traditional statistical methods, generalized linear models (GLMs) and generalized additive models (GAMs) (Martínez-Rincón et al., 2012; Maunder and Punt, 2004). Data mining techniques, such as artificial neural networks (ANNs) (Maunder and Punt, 2004; Hinton and Maunder, 2004), regression trees (RTs) (Norcross et al., 1997; Walsh and Kleiber, 2001), and support vector machine

Submitted on: 1/November/2018
Approved on: 12/November/2019
Associate Editor: Maria Gasalla
Editor: Rubens M. Lopes

(SVM) (Shono, 2014; Li et al., 2015) have also been used. Among those methods, SVM has proven to provide the best performance in CPUE standardization (Li et al., 2015; Shono, 2014; Yang et al., 2015b).

SVM is a classification and regression method based on the principle of structural risk minimization (Li et al., 2015; Shono, 2014; Yang et al., 2015b). It is an effective method to avoid local optima and has unique advantages in dealing with complex problems such as limited samples, high dimensional and nonlinear data. The success of SVM depends on the choice of its kernel parameters and penalty factor, and the key to improve the accuracy is to select the appropriate parameters (Zhou et al., 2019). However, the selection of appropriate SVM hyper-parameters is still challenging for casual users and has not been discussed in CPUE standardization. At present, there are many parameters optimization methods. Yang et al. (2015b) applied Grid Search methods to select the SVM parameters. Nieto et al. (2015) used the particle swarm optimization (PSO) algorithm to optimize SVM parameters, and the results confirm the feasibility and superiority of the proposed optimization method. Zhou et al. (2019) adopted the genetic algorithm (GA) for SVM parameter optimization. These methods have been applied for parameter optimization in other applications but have not been discussed in regard to CPUE standardization.

Bigeye tuna (*Thunnus obesus*) is mainly exploited in the tropical region of the Atlantic Ocean by longliners (Hsu and Lee, 2003; Andrade, 2015). The stock status in the Atlantic and CPUE trend of the stock quickly increased during 1980 and decreased from 1990 to the present. Due to its high economic value and stock status, the species has become one of the most concerned by regional management organizations (Hsu and Lee, 2003; Soto et al., 2009; Chassot et al., 2016; Andrade, 2015). Hsu and Lee (2003) applied GLM to standardize the Taiwanese longline CPUE for bigeye tuna in the Atlantic Ocean and estimate an annual trend. Soto et al. (2009) and Katara et al. (2016) applied GLM standardization for the purse-seine bigeye tuna CPUE. The GLM model has disadvantages when standardizing CPUE, such as the need to specify error distribution assumptions (Li et al., 2015; Shono, 2014).

Bigeye tuna are caught mainly by longline, especially the adult fish. The efficiency of longline gear differs among the depth of hooks and their relationship with the swimming depth of the fish (Matsumoto et al., 2005). Bertrand (2002) verified that the longline catch rate of bigeye tuna is influenced by its vertical habitat. In addition, the stock assessment procedures for bigeye tuna are commonly based on the analysis of data from longline fisheries. The potential vertical habitat of bigeye tuna is essential to the standardization of CPUE indices currently utilized (Ward and Myers, 2005). The thermocline plays a key role to determine the vertical habitat preferences of large tropical pelagic fish (Bertrand et al., 2002; Schaefer and Fuller, 2010; Yang et al., 2015a). Therefore, for analyzing stocks caught by longline, subsurface information seems very important. Nevertheless, most of the studies on CPUE standardization use surface data as environment variables.

In order to evaluate the influence of parameter selection on SVM performance in CPUE standardization, the SVM model was applied to standardize longline fishery catch rates in the Atlantic Ocean and three parameters optimization methods were compared, namely Grid Search and two improved hybrid algorithms, SVMs in combination with the particle swarm optimization (PSO-SVM) and genetic algorithms (GA-SVM).

Argo profile buoy data were adopted to compute the thermocline characteristics used as input environmental variables. The mean absolute error (MAE), mean square error (MSE), correlation coefficients and the normalized mean square error (NMSE) were calculated to evaluate the predictive performance of the different algorithms.

MATERIALS AND METHODS

Study site

Bigeye tuna mainly inhabit tropical and temperate waters worldwide, and the tuna longline fishery occurs mostly in tropical areas. The study area is defined as 60°W – 20°E and 25°S – 25°N (Fig. 1). Bigeye tuna (*Thunnus obesus*) is mainly exploited in this region in the Atlantic Ocean (Hsu and Lee, 2003).

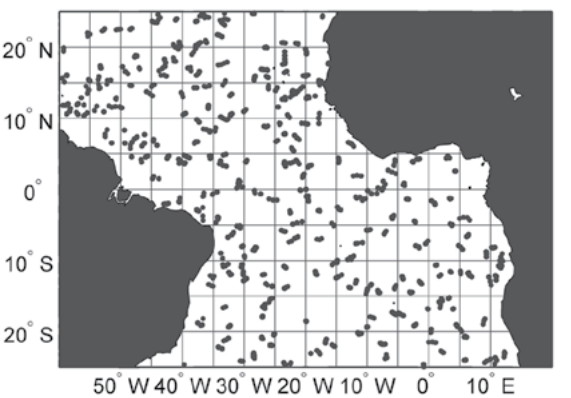

Figure 1. The study area and spatial distribution of Argo profile data in December 2017.

There were 273 distinct Argo buoys released by surrounding countries (United States of America, Canada, Brazil, Europe, Germany, France, Spain, United Kingdom and others) active in this area as in December 2017 (Fig. 1).

Fishery data

Bigeye tuna longline data and Argo buoy data from 2007 to 2017 were used. The longline catch and effort data were compiled from the International Commission for the Conservation of Atlantic Tunas (ICCAT) website. Data were extracted from 2007 to 2017 to match the Argo profile buoys data. Fishing data included number of hooks, fishing time, longitude and latitude and number of bigeye tuna caught. The spatial resolution was 5°×5° and the temporal resolution was a month. Nominal CPUE was defined as the number of individuals caught per 1000 hooks on a 5°×5° grid. The nominal CPUE was calculated by following equation: $\mathrm{CPUE}_{(i,j)} = \frac{N_{\mathrm{fish}(i,j)} \times 1000}{N_{\mathrm{hook}(i,j)}}$. The $\mathrm{CPUE}_{(i,j)}$, $N_{fish(i,j)}$ and $N_{hook(i,j)}$ were the average CPUE, monthly bigeye tuna number and monthly hooks, respectively.

Environmental data

Argo deployments began in 2000, and the array was 100% complete in November 2007. Therefore, in this paper, the Argo buoy data during 2007–2017 downloaded from China Argo real-time data central (http://argo.org.cn/english/) were used to characterize subsurface environments. The Argo buoy data were scattered in the vertical and horizontal direction (Fig. 1). Akima interpolation methods (Akima, 1970) were applied to fit the water temperature data profile in 2-m intervals before estimating the thermocline and subsurface temperatures. Then, the temperature gradient in a vertical direction was estimated using the simple relationship $\Delta t/\Delta h$, where Δt and Δh are the differences in temperature and depth.

According to the method developed by Zhou et al. (2002), 0.05 ºC m^{-1} was adopted as the threshold value to identify the thermocline (the upper and lower boundaries of temperature and depth) using a stepwise discriminant analysis. The details of the computational and determination method can be found in Zhou et al. (2002) and Yang et al. (2015b). If there were several thermocline layers at one point, the upper and lower boundaries of temperature and depth were selected as those of the first and last thermocline layers, respectively.

All the scattered temperature values of the upper and lower boundaries of temperature and all the scattered depth values of the upper and lower boundaries of the thermocline in the horizontal direction were extracted for all years and months and grouped accordingly. Then the contour values with 1º × 1º spatial resolution were calculated using kriging interpolation methods (Yang et al., 2008; 2013). To match with the catch data, the product data was averaged into the 5º×5º spatial resolution.

Support vector machine (SVM)

SVM is a pattern recognition method developed from statistical learning theory (STL) based on the idea of structural risk minimization principle (SRM) (Sun and Zou, 2015). The SVM was originally developed for classification and was later generalized to solve regression problems by importing the insensitive loss function ε (Nieto et al., 2019), and thereby described as support vector regression (SVR).

The basic idea of SVR is briefly described here. The assumption is that there are training data sets $\{(x_i, y_i), i=1,2,...l\}$, where $x_i \in R^d$ and $y_i \in R$. $X=\{x_1, x_2 \cdots x_n\}$are the input data and $Y=\{y_1, y_2 \cdots y_n\}$ is the model output value. The aim of SVR is to find an optimal hyper-plane H. The distance of all the samples to H is small. We wish to predict a real-valued output $f(x)$ for the observed value y. So that the regression functions in high dimensional feature space is $f(x)=w \cdot \varphi(x)+b$. Where "•" is dot product, w is the weight vector and b is the threshold, $\varphi: X \to Z$ is a transformation of the input space into a new space Z, usually a high dimension space. The penalty function is defined as $L = \begin{cases} 0, & |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon, & |y - f(x)| \ge \varepsilon \end{cases}$

Where ε is an insensitive loss function. This problem is transformed into the optimal problem w and b by introducing the relaxation variable ξ and penalty factor c. $\min \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$,

$$s.t. \begin{cases} y_i - w \cdot \varphi(x_i) - b \le \varepsilon + \xi_i^+ \\ -y_i + w \cdot \varphi(x_i) + b \le \varepsilon + \xi_i^-, i = 1,2,\cdots 1. \\ \xi_i^+ \ge 0, \xi_i^- \end{cases}$$

This is the ε-SVM regression. The solution could be obtained from the dual problem. The kernel function is defined as $K(x,x') = (\varphi(x), \varphi(x'))$. The kernel function is useful because many regression problems cannot be linearly regressed in the space of the inputs x, which might be in a higher dimensional feature space given a suitable mapping. At present, there are four commonly used kernel functions (linear, polynomial, gauss and sigmoid) (Zhou et al., 2019).

The RBF has only one parameter and fewer numerical difficulties compared with other functions (Wang et al., 2003). Therefore, the RBF is very suitable for nonlinear and high-dimensional data and was adopted as kernel function in this study. Since the penalty factor *c* and the kernel parameter g need to be manually set and the RBF-SVM parameters directly affect the feature space mapped, the parameter setting is a key step in the method application (Zhou et al., 2019).

Genetic algorithm (GA)

The Genetic Algorithm (GA) is an adaptive heuristic search algorithm designed to mimic the process of genetic selection and natural selection from the biological evolution theory, and realize the evolution of the population through natural selection, crossover and mutation. Genetic algorithms have been used in science and engineering as adaptive algorithms for solving practical problems. After setting the random initial starting point and constructing the fitness function, GA finds the global optimal solution in the search space according the search strategy. A GA pseudo-code is as following (Zhou et al., 2019; Phan et al., 2016).

1: t=0; Initialize $P(t)=\{x_1,x_2,\cdots,x_n\}$, where n is the number of individuals
2: Calculate the fitness of each individual in $P(t)$
3: Breed new generation individual $P''(t)$ through selection, crossover and mutation
4: Calculate the fitness of each individual in $P''(t)$
5: Determine $P(t+1)=P''(t)$ and set t=t+1
6: Return step 3, if the terminating condition is not satisfied

Particle swarm optimization (PSO) algorithm

Swarm optimization (Kennedy and Eberhart, 1995) is an optimization technique based on the metaphor of social behavior. In this algorithm, the particles renew their location by tracking the population best (*Pbest*) and the global best (Gbest). If the current value is better than *Pbest*, the *Pbest* value and location are replaced by that of the current particle. Then, the current value was compared with the Gbest. If the current value came out to be better than *Gbest*, *Gbest* was set equal to the current value of the particle (Nieto et al., 2015; Ghosh et al., 2019).

A possible solution $X_i=(x_{i1}, x_{i2},\cdots,x_{iD})^T$ is called a swarm particle and represents its position in the search space of possible solutions. *D* is the particle dimension. The particle position X_i^0 and its velocity V_i^0 are initially set randomly. The value of the fitness function is then calculated for each particle. The velocity and position of the particle are then changed according to the following equations: $V_{id}^{k+1}=\varpi_1 V_{id}^k+c_1 r_1(Pbest_{id}^k-X_{id}^k)+c_2 r_2(Gbest_d^k-X_{id}^k)$ and $X_{id}^{k+1}=X_{id}^k+V_{id}^k$ where c_1 and c_2 are the acceleration constants. The r_1 and r_2 are random real numbers between 0 and 1. ε_1 is the inertia weight, and is used to control the impact of the history of the velocity on the current one.

To control the diversity of a particle, the mutation algorithm of GA is introduced to renew the particle according to the equation $\begin{cases} X(i,j)=(20-1)*rand+1 & ceil(2*rand)=1 \\ X(i,j)=(\max(g)-\min(g))*rand+\min(g) & ceil(2*rand)=2 \end{cases}$

The PSO pseudo-code is as follows (Nieto et al., 2015):

Input: initialization population of particle Xi = (x_{i1}, x_{i1},···x_{in}), i = 1,2, ··· N;
%N is the number of particles in the population
Output: the best solution g and its corresponding value f_{min}=min($f(x)$)
1:init_particles;
2:eval=0;
3:**while** termination_condition_not meet do
4:**for** i=1:N do
5:f_i=evaluate_the_new_solution(X_i)
6:eval=eval+1;
7:**if** $f_i \le Ibest_i$ **then**
8:$I_i=X_i$;$Ibest_i=f_i$;//save the local best solution
9:**end if**
10:**if** $f_i \le f_{min}$ **then**
11: $g=X_i$, $f_{min}=f_i$;//save the global best solution
12:**end if**
13:X_i=Generate_new_solution(X_i)
14: **end for**
15: **end while**

Grid search algorithms

The first step was to set the minimum and maximum value of the penalty factor c and the kernel parameter g. Second, the values of c and g cycled from $\mathbf{2^{(min\ c)}}$ to $\mathbf{2^{(max\ c)}}$ and $\mathbf{2^{(min\ g)}}$ **to** $\mathbf{2^{(max\ g)}}$ in a two-layer cycle. K-fold-cross-validation algorithms (Arlot and Celisse, 2010) was used on the training data in each step and then the average square error (MSE) of the remaining k-1 subsamples was calculated. At last, the output values of c and g corresponding to the minimum MSE of the two-layer cycles were chosen as the best model parameters. In this

paper, the parameters c and g ranged from 2^{-10} **to** 2^{10} and the interval was set to 1. The k value of cross validation was set to 5.

A pure pseudo-code of the Grid search follows:

```
Start
    bestAccuracy=0;
    bestc=0;
    bestg=0;
    for c=2^(min c): 2^(max c)
        for g=2^(min g) : 2^(max g)
            Division the training set into 5, train(1), train(2),
train(3), train(4), train(5). Training model by leave-one-
out cross validation and calculate mean error cv.
            If cv>bestAccuracy
            best Accuracy = cv; best c = c; best g = g;
            end
        end
    end
over
```

GA-SVM

The genetic algorithm was implemented by using the Sheffield toolbox (Holland, 1992). The population of the first generation was randomly selected. Each individual member in the population (e.g. c and g) was coded using a binary string with a length of 20. The number of the binary string of each individual was assumed to be equal to the number of the descriptors (e.g., 2). Therefore, the length of a "chromosome" in the population was 40. The main operators were crossover, selection, reinserts and mutation. First, the upper and lower bounds of the two SVM parameters c and g were specified [0.1, 100]. The application probability of these operators was 0.7 for crossover and 0.035 for mutation, respectively. The selection method was a random ergodic process and the selection probability was 0.9. The population size was 20. If a fitness criterion (10^{-2}) or a maximum number of iterations (200) was reached, the computation would stop. A 5-fold cross-validation algorithm on the training data for each individual was used, with the MSE of the remaining k-1 subsamples as the fitness value for identifying suitable parameters for the SVM model. The best parameters were applied to construct the SVM. Fig. 2 shows the flowchart of the GA-SVM model.

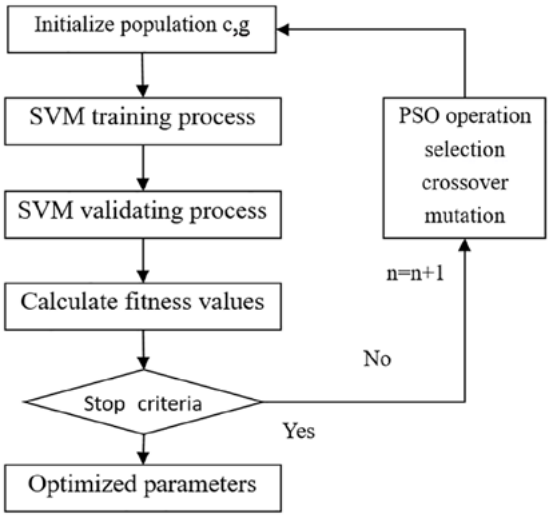

Figure 2. Flowchart of the hybrid GA-SVM model.

PSO-SVM

The upper and lower bounds of the SVM parameters c and g were specified as [0.1, 100]. The values for the two SVM parameters were then generated randomly within the specified bounds for each particle. The population size was set to 20. Then these parameters were fed into a SVM model, and 5-fold cross-validation algorithms were used to evaluate model performance. The fitness value adopt as mean square error $\frac{1}{n}g\sum_{i=1}^{n}(y_i - f_i)^2$ where y_i is observed value and f_i is predicted value. Fig. 3 shows the flowchart of PSO-SVM model. The best parameters were applied to construct the SVM.

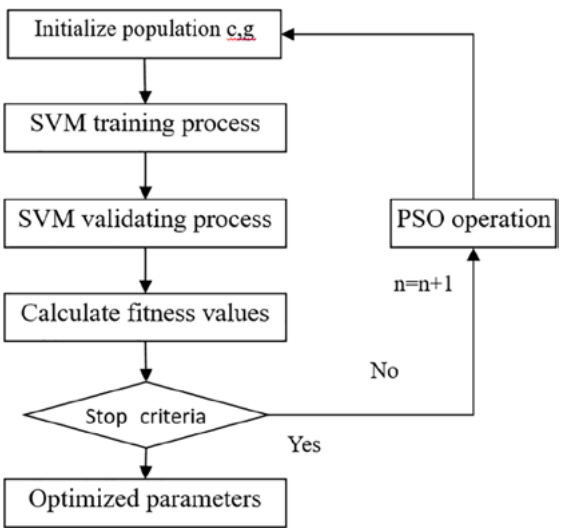

Figure 3. Flowchart of the hybrid PSO-SVM model.

Data Preprocessing

The input variables included year, month, latitude, longitude and four thermocline factors: the upper depth boundary of the thermocline (UDBT), the lower depth boundary of the thermocline (LDBT), the upper temperature boundary of the thermocline (UTBT), and the lower temperature boundary of the thermocline (LTBT). The nominal CPUE was defined as the continuous response variable in all three SVM models. There are differences in the values of different variables in the original data. Therefore, the data were normalized according to the equation $y=(x-x_{min})/(x_{max}-x_{min})$.

All the records were randomly divided into two subsets: the training data set (80% of the total) and the testing set (20% of the total). The training data set was used to find the model parameters and the testing set was used to predict the model.

Model evaluation and year trend

The goodness of fit of the models were tested by the MAE, MSE, and three correlation coefficients (Pearson's, Kendall's and Spearman's coefficients of correlation) and the NMSE according to the equation $NMSE=\Sigma(y-\hat{y})^2/\Sigma(y-\bar{y})^2$, where y was the predicted value and $\hat{y}$ represent the nominal CPUE value, respectively. $\bar{y}$ is the average value of the nominal CPUE.

A generalized additive model (GAM) was constructed based on all data to examine the nature of the relationship between the nominal CPUE and the environmental variables. Following Maury et al. (2001), we assumed a normal distribution for log (CPUE + 1). Year, month, latitude, longitude, and four thermocline factors (UDBT, LDBT, UTBT, LTBT) were selected as input variables. All explanatory variables were modeled as a spline function (df=4). In order to compare with SVM, MSE, MAE, three correlation coefficients and NMSE were calculated.

After calculating the predicted CPUE value at each observed point, average values of all predicted CPUEs in each year were computed and mapped.

All simulations and analyses were carried out with Matlab 2010b 32-bit software, using the packages "libsvm-mat-2.89-3" for SVM and Sheffield toolbox for GA. The PSO algorithm was constructed on Matlab 2010 software without toolbox. The GAM was constructed in the R programming environment using the gam function in the mgcv package (Wood, 2006). Model selection was performed manually, and we retained significant candidate predictors, minimized the Akaike information criteria (AIC) and increased the amount of explained deviance.

RESULTS

Table 1 shows the values of *c* and *g* derived from the three models. The values of the MSE and MAE in the training, testing data and dataset were displayed in Table 2. Table 3 presents the values of Pearson's, Kendall's, and spearman's correlation coefficients. Table 4 expresses the values of NMSE.

Table 1. The parameters of c and g of different models.

	c	g
Grid search	5.66	64
GA-SVM	5.77	50
PSO-SVM	7.12	60.25

Table 2. Comparison of error between different models.

Model	MSE			MAE		
	Training data	Testing data	All data	Training data	Testing data	All data
Grid search	3.62	**8.64***	4.62	0.68	**1.81***	**0.90***
GA-SVM	0.57	**12.37**	2.94	0.51	**2.11**	**0.83**
PSO-SVM	0.4*	**10.92**	2.5*	0.46*	**2.02**	**0.77***

Table 3. Comparison of correlation coefficients between different models.

Model	Person's			Kendall's rank			Spearman's rank		
	Training data	Testing data	All data	Training data	Testing data	All data	Training data	Testing data	All data
Grid search	0.88	**0.69***	0.85	0.85	**0.55***	0.77	0.95	**0.73***	0.9
GA-SVM	0.99	**0.6**	0.91	0.92	**0.52**	0.83	0.98	**0.65**	0.92
PSO-SVM	0.99*	**0.61**	0.92*	0.94*	**0.52**	0.84*	0.99*	**0.67**	0.92*

The PSO-SVM and GA-SVM had particularly high performances of indicative values for the training data and dataset, but the MSE, MAE, correlation coefficients and NMSE in the testing dataset were not as good as that for the training dataset. The performance of the Grid search was better than that of the PSO and GA algorithms when comparing the results of the testing (Tables 2, 3 and 4). Regarding the comparison of PSO-SVM and GA-SVM, the goodness of PSO-SVM is marginally better than GA-SVM by different indicative values in the training data and testing dataset (Tables 2, 3 and 4).

The annual trends in standardized CPUE obtained from the three models and nominal CPUE were similar (Fig. 4). The average CPUE for all three algorithms was almost always lower than that of the nominal CPUE values. The trends in bigeye tuna relative abundance, estimated with the three models, fluctuated from 2007 to 2017. The nominal CPUE trends were also highly correlated with the various standardized trends. Overall, the bigeye tuna relative abundance was low from 2007 to 2011, increased during 2011-2013, then decreased from 2015 to 2017.

The results of the GAM analysis indicated that all predictor variables were retained at the 0.001 level (Table 5). The GAMs explained 29.6% of the null deviance (Table 6). The MSE, MAE, correlation coefficients and NMSE were inferior to SVM model. The annual trend in standardized CPUE obtained from the GAM was lower than SVM and nominal CPUE (Fig. 4). The general relationships between *ln(CPUE+1)* and all variables are shown in Fig. 5. The hooking rates were low before 2012 and then increased; seasonally, the hooking rates decreased from January to August, and then increased. The nominal CPUE values increased from south to north. The CPUE value was highest at approximately 40°W. UTBT, LDBT, and LTBT had a negative effect. The general relationship between nominal CPUE and UDBT was characterized by a convex shape.

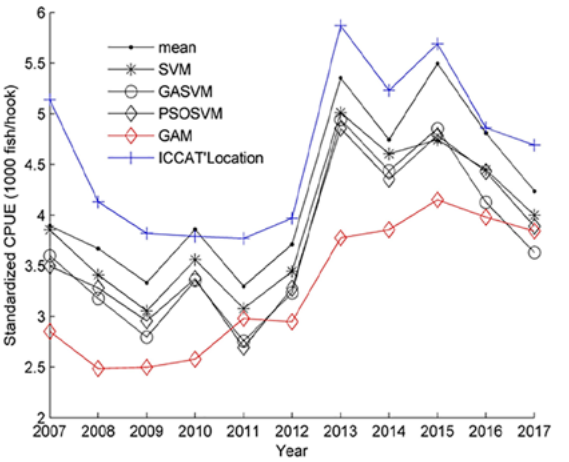

Figure 4. Comparisons of estimates of annual standardized CPUE and the nominal CPUE (mean).

Table 4. The value of NMSE between different models.

Model	NMSE		
	Training data	Testing data	All data
Grid search	0.24	0.53*	0.3
GA-SVM	0.04	0.75	0.2
PSO-SVM	0.03*	0.730	0.16*

Table 5. F-test for the significance of nonparametric effects.

Variable	*d.f.*	*F*	*p*	Variable	*d.f.*	*F*	*p*
s(year)	3	59.21	<0.000001	*s(UTBT)*	2.99	64.82	<0.000001
s(month)	2	9.68	0.00032	*s(UDBT)*	2.86	51.33	<0.000001
s(lat)	3	96.23	< 2e-16	*s(LTBT)*	2	84.98	<0.000001
s(lon)	3	46.65	<0.000001	*s(LDBT)*	2.45	66.63	<0.000001

Table 6. Statistical characteristics of the GAM model.

Deviance explained (%)	AIC	R^2 (adj)	MSE	MAE	Person's	Kendall's rank	Spearman's rank	NMSE
29.6	11381.74	0.293	13.38	2.33	0.44	0.37	0.52	0.86

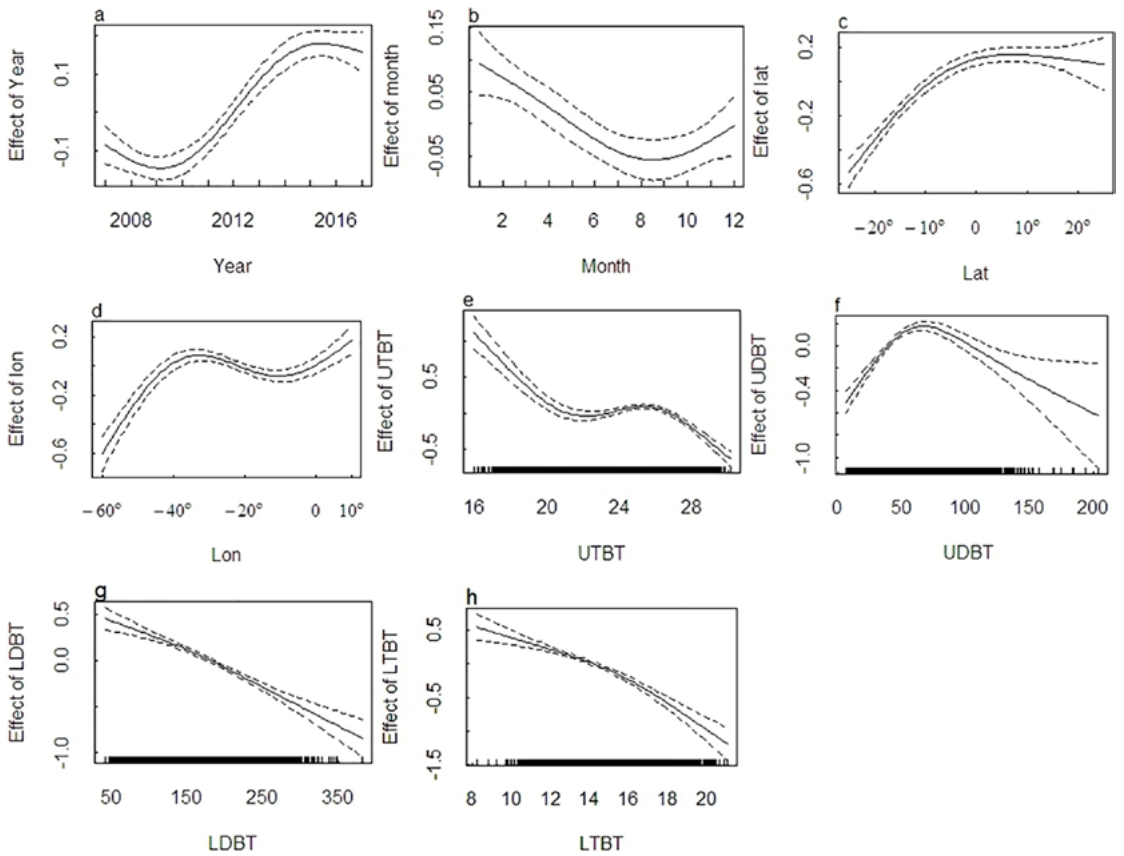

Figure 5. The GAM-derived effects of eight predictors on CPUE.

DISCUSSION

SVM is widely applied for classification and regression problems for its high accuracy and ability to deal with high-dimensional data. SVM models have been applied in CPUE standardization recently and showed higher performances than traditional standardization algorithms, such as ANNs, RF, GAM and GLM (Li et al., 2015; Shono, 2014; Yang et al., 2015b). Here the SVM model applied to standardized longline fishery catch rates in the Atlantic Ocean appeared to be more reliable compared to GAM. The average CPUE for all three optimization algorithms was always lower than that of the nominal CPUE values, which coincide with previous results (Li et al., 2015). The annual trends were different from previous work due to data series differences (Hsu and Lee, 2003; Andrade, 2015). The data used by Hsu and Lee (2003) and Andrade (2015) were up to the year 2010. The trends in CPUE suggest that the time series could be divided into two periods by 2012. The catch rates after 2012 are higher than before 2012. This may relate to the resource management in Atlantic tuna in recent years.

The year trend of standardized and nominal CPUE was similar with the standardized index computed in the Report of the 2018 ICCAT bigeye tuna stock assessment meeting (ICCAT tropical tuna working group, 2018) (Fig. 4). The CPUE decreased from 2015. Decreasing trends suggesting more management work should be considered to maintain the stock status. The ICCAT trend is smoother that the SVM and nominal trends during 2010 to 2011. The year trend of SVM and nominal CPUE was always lower than the ICCAT. The difference may have resulted from the data resolution and study area. The monthly 5×5 geographical grid in the study area was used only in this paper.

The SVM high accuracy has been reported in previous research (Li et al., 2015; Shono, 2014; Yang et al, 2015b). Nevertheless, SVM has disadvantages, like other data mining models. For instance, it is difficult to estimate the confidence intervals of CPUE trends and inspect the effect of input variables on the CPUE. GAM performances

are not as efficient as SVM but allows to examine the nature of the relationship between the nominal CPUE and spatial, temporal and environmental variables. (Fig. 5). The effects of the eight variables on nominal CPUE were nonlinear, except LDBT and LTBT.

Analyses based on depth-specific catch rates can lead to serious misinterpretation of abundance trends (Bigelow and Maunder, 2007). Considering the tuna depth distribution can effectively reduce the CPUE uncertainty (Prince et al., 2015). There are strong relationships between bigeye tuna depth and CPUE (Abascal et al., 2018) and the thermocline is a key factor in depth distribution of bigeye tuna (Bertrand et al., 2002; Houssard et al., 2017; Abascal et al., 2018). All thermocline parameters were significant in the GAM model support, so that the thermocline was adopted as an environmental factor in CPUE data standardization. In fact, the thermocline is more important than the sea surface temperature (SST). Bigeye tuna descend to a depth well below the thermocline to prey on the small nektonic organisms of the deep scattering layer (DSL) after dawn (Evans et al., 2008; Schaefer and Fuller, 2010). SST is not the main determinant of the tuna areal distribution and relative abundance in the equatorial Atlantic (Zagaglia et al., 2004). For example, bigeye tuna stayed preferentially in zones where the sea surface temperature (SST) was higher than 26°C. However, high hooking rates were observed off Namibia, where the SST was lower than 20°C (Yang et al., 2013). Conversely, small individuals of bigeye tuna were caught in the north of the slope that starts from 15°N of West Africa and ends at 10°N of South America, where the SST was higher than 26 °C (Yang et al., 2013).

The values of *c* and *g* were widely different when derived from three different algorithms. Under different parameters conditions, the model performances were different. The GA and PSO algorithms searched parameters and quickly achieved the best solution. The parameters values (c, g) found by the grid search is in the search space of GA and PSO. If the initial populations of GA or the PSO algorithms contain the parameters values found by the grid search, the GA or the PSO algorithms may obtain the best solution more quickly.

Results suggested that the optimization algorithm of PSO and GA had excellent capability in global spatial searching in the training data set. But the forecasting performances of these two optimization algorithms were not as good as in the training dataset and testing dataset and they were inferior to the Grid search algorithms. Over-fitting is a phenomenon that occurs when the performance error of the model is observed to be very small during training but high in validation. The above results suggested that the GA-SVM and PSO-SVM models may suffer from an over-fitting problem. But in CPUE standardization, models were not used in prediction. In general, considering the indicative values obtained from training data and dataset, the PSO-SVM performed best, and can be applied as an alternative to standardize bigeye tuna in the Atlantic Ocean. The PSO is appropriate to optimize the SVM model parameters.

CONCLUSIONS

In this paper, we applied SVM to standardize bigeye tuna longline fishery catch rates in the Atlantic Ocean and the thermocline was adopted as an environmental factor in CPUE standardization for the first time. In order to increase SVM accuracy, different optimization methods were evaluated. All models performed well. PSO and GA exhibited excellent capability in global spatial searching, however they may suffer from over-fitting problems. PSO had best indicative values and is suggested as an effective method to optimize the SVM parameters.

ACKNOWLEDGEMENTS

This study was supported by the Chinese National Natural Science Foundation (41606138), the Special Funds of Basic Research of Central Public Welfare Institute (2019T09, 2019HY-XKQ03), the Natural Science Foundation of Shanghai (Z53201719870) and the Key Laboratory of Sustainable Exploitation of Oceanic Fisheries Resources (Shanghai Ocean University), Ministry of Education: A1-2006-00-301109.

REFERENCES

ABASCAL, F. J., PEATMAN, T., LEROY, B., NICOL, S., SCHAEFER, K., FULLER, D. W. & HAMPTON, J. 2018. Spatiotemporal variability in bigeye vertical distribution in the Pacific Ocean. *Fisheries Research,* 204, 371-379.

ANDRADE, H. A. 2015. Sensitivity analysis of catch-per-unit-effort of Atlantic bigeye tuna (*Thunnus obesus*) data series applied to production model. *Latin American Journal of Aquatic Research,* 43, 146-161.

ARLOT, S. & CELISSE, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys,* 4, 40-79.

BERTRAND, A., JOSSE, E., BACH, P., GROS, P. & DAGRON, L. 2002. Hydrological and trophic characteristics of tuna habitat: consequences on tuna distribution and longline catchability. *Canadian Journal of Fisheries and Aquatic Sciences,* 59, 1002-1013.

BIGELOW, K. A., BOGGS, C. H. & HE, X. 1999. Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery. *Fisheries Oceanography,* 8, 178-198.

BIGELOW, K. A. & MAUNDER, M. N. 2007. Does habitat or depth influence catch rates of pelagic species? *Canadian Journal of Fisheries and Aquatic Sciences,* 64, 1581-1594.

EVANS, K., LANGLEY, A., CLEAR, N. P., WILLIAMS, P., PATTERSON, T., SIBERT, J., HAMPTON, J. & GUNN, J. S. 2008. Behaviour and habitat preferences of bigeye tuna (*Thunnus obesus*) and their influence on longline fishery catches in the western Coral Sea. *Canadian Journal of Fisheries and Aquatic Sciences,* 65, 2427-2443.

GARCIA, S. M. & ROSENBERG, A. A. 2010. Food security and marine capture fisheries: characteristics, trends, drivers and future perspectives. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 365, 2869-2880.

GHOSH, G., MANDAL, P. & MONDAL, S. C. 2019. Modeling and optimization of surface roughness in keyway milling using ANN, genetic algorithm, and particle swarm optimization. *The International Journal of Advanced Manufacturing Technology,* 100, 1233-1242.

GUAN, W. J., TIAN, S. Q., WANG, X. F., ZHU, J. F. & CHEN, X. J. 2014. A review of methods and model selection for standardizing CPUEs. *Journal of Fishery Sciences of China,* 21, 852-862.

HINTON, M. G. & MAUNDER, M. N. 2004. Methods for standardizing CPUE and how to select among them. *Collective Volume of Scientific Papers ICCAT,* 56, 169-177.

HOLLAND, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* Cambridge, MIT Press.

HOUSSARD, P. A., LORRAIN, A., TREMBLAY-BOYER, L., ALLAIN, V., GRAHAM, B. S., MENKES, C. E., PETHYBRIDGE, H., COUTURIER, L. I. E., POINT, D., LEROY, B., RECEVEUR, A., HUNT, B. V., VOUREY, E., BONNET, S., RODIER, M., RAIMBAULT, P., FEUNTEUN, E., KUHNERT, P. M., MUNARON, J. M., LEBRETON, B., OTAKE, T. & LETOURNEUR, Y. 2017. Trophic position increases with thermocline depth in yellowfin and bigeye tuna across the Western and Central Pacific Ocean. *Progress in Oceanography,* 154, 49-63.

HSU, C. & LEE, H. 2003. General linear mixed model analysis for standardization of Taiwanese longline CPUE for bigeye tuna in the Atlantic Ocean. *Collective Volume of Scientific Papers ICCAT,* 55, 1892-1915.

ICCAT (Tropical tuna Working group). 2018. *Report of the 2018 ICCAT bigeye tuna stock assessment meeting* [Internet]. Available from: https://www.iccat.int/Documents/Meetings/Docs/2018/REPORTS/2018_BET_SA_ENG.pdf

KATARA, I., GAERTNER, D., MAUFROY, A. & CHASSOT, E. 2016. Standardization of catch rates for the eastern tropical Atlantic bigeye tuna caught by the French purse seine DFAD fishery. *Collective Volume of Scientific Papers ICCAT,* 72, 406-414.

KENNEDY, J. & EBERHART, R. C. 1995. Particle Swarm Optimization. *Proceedings of IEEEE International Conference on Neural Networks,* Perth, Australia, 1942-1948.

LI, Z., YE, Z., WAN, R. & ZHANG, C. 2015. Model selection between traditional and popular methods for standardizing catch rates of target species: A case study of Japanese Spanish mackerel in the gillnet fishery. *Fisheries Research,* 161, 312-319.

MARTÍNEZ-RINCÓN, R. O., ORTEGA-GARCÍA, S. & VACA-RODRÍGUEZ, J. G. 2012. Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*Acanthocybium solandri*) in the Mexican tuna purse-seine fishery. *Ecological Modelling,* 233, 20-25.

MATSUMOTO, T., SAITO, H. & MIYABE, N. 2005. Swimming behavior of adult bigeye tuna using pop-up tags in the central Atlantic Ocean. *Collective Volume of Scientific Papers ICCAT,* 57, 151-170.

MAUNDER, M. N. & LANGLEY, A. D. 2004. Integrating the standardization of catch-per-unit-of-effort into stock assessment models: testing a population dynamics model and using multiple data types. *Fisheries Research,* 70, 389-395.

MAUNDER, M. N. & PUNT, A. E. 2004. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research,* 70, 141-159.

NIETO, P. J. G., FERNÁNDEZ, A. J. R., SUÁREZ, V. M. G., MUÑIZ, C. D., GARCIA-GONZALO, E. & BAYÓN, R. M. 2015. A hybrid PSO optimized SVM-based method for predicting of the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir: A case study in Northern Spain. *Applied Mathematics and Computation,* 260, 170-187.

NORCROSS, B. L., MÜTER, F. J. & HOLLADAY, B. A. 1997. Habitat models for juvenile pleuronectids around Kodiak Island, Alaska. *Fishery Bulletin,* 95, 504-520.

PAULY, D., HILBORN, R. & BRANCE, T. A. 2013. Fisheries: Does catch reflect abundance? *Nature,* 494, 303-306.

PHAN, A.V., NGUYEN, M. L. & BUI, L. T. 2017. Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems. *Applied Intelligence,* 46, 455-469.

PRINCE, E. D., DEWARH, H., HOOLIHAN, J. P., LUO, J., DIE, D., MAUNDER, M. & GOODYEAR, P. 2015. *Incorporating hypoxia-based habitat compression impacts into the stock assessment process for tropical pelagic billfish and tuna*. NOAA Fisheries Proposal: 2015 Habitat Assessment Improvement Project (HAIP), Silver Springs, NOAA.

RICKER, W. E. 1975. Computation and interpretation of biological statistics of fish populations. *Bulletin of the Fisheries Research Board of Canada,* 191, 1-382.

SCHAEFER, K. M. & FULLER, D. W. 2010. Vertical movements, behavior, and habitat of bigeye tuna (*Thunnus obesus*) in the equatorial eastern Pacific Ocean, ascertained from archival tag data. *Marine Biology,* 157, 2625-2642.

SHONO, H. 2014. Application of support vector regression to CPUE analysis for southern bluefin tuna *Thunnus maccoyii* and its comparison with conventional methods. *Fish Science,* 80, 879-886.

SOTO, M., PALLARÉS, P., MOLINA, A. D. D. & GAERTNER, D. 2009. Standardized CPUE for juvenile yellowfin, skipjack and bigeye tuna from the european purse seine fleet in the Atlantic Ocean from 1991 to 2006. *ICCAT Collective Volume of Scientific Papers,* 64, 1044-1053.

SUN, X. Q. & ZOU, L. Y. 2015. Research on application of support vector machine in fault recognition of bearings. *Process Autom Instrum,* 36,12-15.

WALSH, W. A. & KLEIBER, P. 2001. Generalized additive model and regression tree analyses of blue shark (*Prionace glauca*) catch rates by the Hawaii-based commercial longline fishery. *Fisheries Research,* 53, 115-131.

WANG, W., XU, Z., LU, W. & ZHANG, X. 2003. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing,* 55, 643-663.

WARD, H. G., ASKEY, P. J., POST, J. R. & ROSE, K. 2013. A mechanistic understanding of hyperstability in catch per unit effort and density-dependent catchability in a multistock recreational fishery. *Canadian Journal of Fisheries and Aquatic Sciences,*70, 1542-1550.

WARD, P. & MYERS, R. A. 2005. Inferring the depth distribution of catchability for pelagic fishes and correcting for variations in the depth of longline fishing gear. *Canadian Journal of Fisheries and Aquatic Sciences,* 62,1130-1142.

YANG, S. L., ZHANG, Y., FAN, W. & DAI, Y. 2012. Relationship between the temporal-spatial distribution of fish in bigeye tuna fishing grounds and the thermocline characteristics in the tropical Indian Ocean. *Journal of Fishery Sciences of China,* 19, 679-689.

YANG, S. L., MA, J. J., WU, Y. M. & ZHOU, W. F. 2008. Study on the reconstruction of Pacific temperature arena with Argo data based on the Kriging methods. *Marine Fisheries,* 30, 13-18.

YANG, S. L., ZHANG, S. M., JIANG, X. W., ZOU, B., HUA, C. J. & ZHOU, W. F. 2013. Seasonal variability of thermocline in *Thunnus obesus* and *Thunnus albacares* fishing ground in the Tropic Atlantic Ocean. *Journal of Applied Oceanography,* 32, 349-357.

YANG, S., MA, J. J., WU, Y. M., FAN, X. M., JIN, S. F. & CHEN, X. Z. 2015a. Relationship between temporal–spatial distribution of fishing grounds of bigeye tuna (*Thunnus obesus)* and thermocline characteristics in the Atlantic Ocean. *Acta Ecologica Sinica,* 35, 1-9.

YANG, S., ZHANG, Y., ZHANG, H. G. & FAN, W. 2015b. Comparison and analysis of different model algorithms for CPUE standardization in fishery. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE),* 31, 259-264.

ZAGAGLIA, C. R., LORENZZETTI, J. A. & STECH, J. L. 2004. Remote sensing data and longline catches of yellowfin tuna (*Thunnus albacares*) in the equatorial Atlantic. *Remote Sensing of Environment,* 93, 267-281.

ZHOU, T., LU, H. L., WANG, W. W. & YONG, X. 2019. GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing,* 75, 323-332.

ZHOU, Y. X., LI, B. L., ZHANG, Y. J. & BA, L. C. 2002. World oceanic thermocline characteristics in winter and summer. *Marine Science Bulletin,* 21, 16-22.