A Model for Fishery Forecast Based on Cluster Analysis and Nonlinear Regression

H.C. Yuan, M.X. Tan, Y.T. Gu
Shanghai Ocean University
Shanghai, China

Y. Chen
School of Computing and Information Systems
University of Tasmania Hobart
Tasmania, Australia

Abstract—**There has been an increasing amount of research in the relationship between environmental factors and fishing yield. This paper adds to the body of knowledge by developing a new model for forecasting fishing yield. The model combines fishery domain expert knowledge, marine environmental factor data such as water temperature, chlorophyll concentration and sea surface level as base data and applies cluster analysis that incorporates function fitting and nonlinear regression for data analysis and processing. The model is tested for forecast accuracy and the test result is compared with those using RBF and SVM, the two methods commonly used for similar purposes. The comparison result reveals this new model increases both the accuracy in fishery forecast and the reliability in guiding fishery production and related activities. It can also help explore and discover the distribution of fishing grounds.**

Keywords-pelagic fishing; fisheries forecasting; cluster analysis; nonlinear regression

I. INTRODUCTION

In the past twenty years, China has experienced steady development in its pelagic fishery. By the end of 2012, the total number of offshore fishing vessels in China reached 2200. Offshore fishing has improved the piscatorial income of the nation and the living standards of its people with an increased supply of aquatic products. It has also contributed to the nation's fishing industry, protected its inshore fishery resources and promoted the development of other related industries [1]. Illex argentinus takes a very important position in the fishery economy regardless of its output value or yield. Research in Illex argentines covers various aspects [2-6]. Bakun and Csirke[7] studied the impact of marine environmental changes on Illex supplement group resources. Song and Xiong[8] reported the correlation between Illex production distribution and sea surface temperature. In spite of the large population in China, existing data still shows China lagging way behind most developed countries in terms of the amount of offshore fishing per capita share. Even though aquatic operators have accumulated an enormous marine fish catch database over the years yet the data has not been used due to lack of data analysis tools. Consequently, aquatic workers are still relying purely on personal knowledge and experience for tasks such as the delineation of fishing waters and the ascertaining of fishing time, resulting in aimless fishing and an overly high cost of fishing overall. There is an urgent need for a sophisticated and reliable analysis tool to process and analyze the fishery data so that it can assist fishery prediction and support more cost effective and efficient fishing [9-11]. It can also stabilize the further development of China's offshore fishing, enhance China's current status in international fishery and reduce the pressure of China's inshore fishery resources.

II. MATERIALS AND METHODS

A. Research Data Sources

For the purpose of this research, the environmental data of Illex argentinus' fishing include sea surface level, sea surface temperature and chlorophyll concentration in Southwest Atlantic in January-April 2000. This data is downloaded from OceanWatch (http://oceanwatch.pifsc.noaa.gov/las/). The corresponding fishery yield or catch data is provided by Shanghai Ocean University.

B. Data analysis Methods and Principles of Clustering

The proposed data analysis method is hierarchical clustering which works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. However a pure hierarchical clustering method suffers from its inability to be adjusted if a merge or split decision turns out to be a poor choice, ie, the method does not allow backtracking for any necessary corrections. Recent studies have emphasized on the integration of hierarchical agglomeration with iterative relocation methods.

Four widely used measures for distance between clusters are presented here. Where $|p-p'|$ is the distance between two objects or points, p and p'; m_i is the mean for cluster, C_i; and n_i is the number of objects in C_i.

Minimum distance: $d_{\min}(C_i, C_j) = \min_{p\in C_i, p'\in C_j} |p-p'|$ (1)

Maximum distance: $d_{\max}(C_i, C_j) = \max_{p\in C_i, p'\in C_j} |p-p'|$ (2)

Mean distance: $d_{mean}(C_i, C_j) = |m_i - m_j|$ (3)

Average distance: $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p\in C_i} \sum_{p'\in C_j} |p-p'|$ (4)

When an algorithm uses the minimum distance, $d_{\min}(C_i, C_j)$, to measure the distance between clusters, it is sometimes called a nearest-neighbor clustering algorithm. Moreover, if the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a single-linkage algorithm. If we view the data points as nodes of a graph, with edges forming a path between the nodes in a cluster, then the merging of two clusters, C_i and C_j, corresponds to adding an edge between the nearest pair of nodes in C_i and C_j. Because edges linking clusters always go between distinct clusters, the resulting graph will generate a tree. Thus, an agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a minimal spanning tree algorithm.

C. Data processing

The data of sea surface level, sea surface temperature, chlorophyll concentration and fishing yield is normalized in this analysis. The formula of normalization is defined in (5) to (9).

$$x'_{year} = \frac{x_{year} - 1995}{2050 - 1995} \tag{5}$$

$$x'_{month} = \frac{x_{month}}{12} \tag{6}$$

$$x'_{latitude} = \frac{x_{latitude} + 90}{180} \tag{7}$$

$$x'_{longitude} = \frac{x_{longitude} + 180}{360} \tag{8}$$

$$x'_i = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}} \tag{9}$$

Where x_i is the sea surface level, sea surface temperature and chlorophyll concentration in $x'_i = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}}$, $x_{i\max}$ represents the maximum value while $x_{i\min}$ represents the minimum value of the month.

A dendrogram of sea surface level, sea surface temperature and chlorophyll concentration data has been created by clustering analysis with MATLAB 7.10.

With the clustering analysis in MATLAB, it is revealed that there are five sets of data which are divided into two categories, so that in the range of non-central fishing ground these minimal sets of data points which are relative to the overall data are outliers.

D. Fitting and Regression Analysis

Since there is no existing reference for quantitative analysis between fishing yield and environmental factors or functions of the exact relationship, data and fitting function are considered in this paper.

Equations (10), (11) and (12) have been created by function fitting of sea surface temperature, sea surface level or chlorophyll concentration with fishing yield data using a software package called 1stOpt after removing outliers.

$$A = y = \sqrt{\left((0.0413 - 0.0726 * x) \big/ 1 - 3.421 * x + 2.940 * x^2 \right)} \tag{10}$$

$$B = y = \frac{1}{(31288.2 + 13527.9 * e^x + 83204.8 \big/ Ln(x))} \tag{11}$$

$$C = y = 0.0336 * e^{-62141 * x} + 0.3380 * e^{-\left((x-0.022)^2 \big/ (-0.001)^2\right)} + 0.0002 * e^{-\left((x-0.1989)^2 \big/ 0.006^2\right)} \tag{12}$$

Where Equation (10) is the function of the relationship between fishing yield and sea surface temperature, Equation (11) is the function of the relationship between yield and sea surface level, and Equation (12) is the function of the relationship between yield and chlorophyll concentration.

The complexity of the model can be reduced with the assumption that these three marine environmental factors are independent of each other. As a result a regression model of the yield (y) and sea surface temperature, x_1, sea surface level, x_2, and chlorophyll concentration, x_3, can be created through the fitting function of the production and each marine environmental factors. See (13) below.

$$y = a * A + b * B + c * C + d \tag{13}$$

Where a, b, c, d are coefficients which are unknown.

The final non-linear regression model (14) is obtained with a, b, c, d and using formula (13):

$$\begin{aligned} y = {} & 0.4539 * \sqrt{\left((0.0413 - 0.0726 * x_1) \big/ 1 - 3.421 * x_1 + 2.940 * x_1^2 \right)} \\ & + 0.8528 * \frac{1}{(31288.2 + 13527.9 * e^{x_2} + 83204.8 \big/ Ln(x_2))} \\ & + 0.8067 * (0.0336 * e^{-62141 * x_3} + 0.3380 * e^{-\left((x_3-0.022)^2 \big/ (-0.001)^2\right)} + 0.0002 * e^{-\left((x_3-0.1989)^2 \big/ 0.006^2\right)}) \\ & - 0.006 \end{aligned} \tag{14}$$

Finally the actual data and the resulting function point chart is shown in Fig. 1.

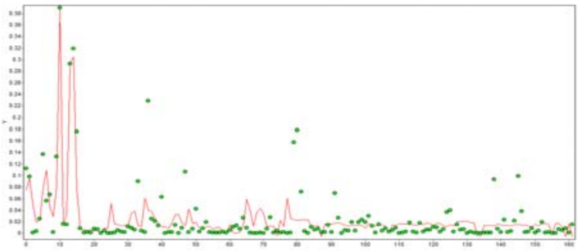

FIGURE I. DIAGRAM OF RELATIONSHIP BETWEEN DATA AND MODEL

III. RESULTS AND DISCUSSION

RBF and SVM are two commonly used methods for fishing forecast. It is worth comparing these two methods with the Nonlinear Regression Model (NRM) proposed in this paper. A few sets of environmental data are randomly selected to forecast fishing production using NRM and the resulting forecast data is compared with the actual measured yield data from the existing database hosted in Shanghai Ocean University. Table. 1 shows the comparison result. A similar process is adopted applying the RBF and SVM methods and the results are shown in Table. 2 and Table. 3.

TABLE I. COMPARISON BETWEEN ACTUAL YIELD AND PREDICTED YIELD USING NRM

Measured Production	Forecast Production	Relative Error
0.1126	0.0752861	0.3314
0.098	0.0949515	0.0311
0.0247	0.023498	0.0487
0.0671	0.0459562	0.3151
0.3898	0.389551	0.0006
0.016	0.0134827	0.1573
0.2923	0.2948493	0.0087
0.3192	0.3039031	0.0479
0.009	0.0069658	0.226
0.0075	0.0048264	0.3565
0.0068	0.0050396	0.2589
0.0122	0.0126166	0.0341
0.0208	0.0247357	0.1892
0.0137	0.0169747	0.239
0.0144	0.0170137	0.1815
0.0269	0.0185812	0.3092
0.0195	0.0143232	0.2655
0.0192	0.0141467	0.2632
0.013	0.0144543	0.1119
0.0156	0.0153768	0.0143
0.0085	0.0113611	0.3366
0.0182	0.0155487	0.1457
0.018	0.0170004	0.0555
0.011	0.013349	0.2135
0.0153	0.0199268	0.3024

TABLE II. COMPARISON BETWEEN ACTUAL YIELD AND PREDICTED YIELD USING RBF

Measured Production	Forecast Production	Relative Error
0.7859	0.467506	0.405133
0.7331	0.472325	0.355715
0.265	0.468734	0.768808
0.5091	0.462923	0.090703
0.4721	0.480593	0.01799
0.8751	0.497989	0.430935
0.57	0.494183	0.133012
0.6179	0.487046	0.211772
0.4422	0.411414	0.06962
0.5917	0.403983	0.31725
0.9767	0.408605	0.581647
0.8226	0.409219	0.50253
0.4681	0.404942	0.134924
0.5009	0.414117	0.173254
0.716	0.42894	0.400922
0.751	0.433081	0.423328
0.6314	0.4287	0.321033
0.434	0.427575	0.014804
0.4189	0.426827	0.018923
0.3603	0.427067	0.185309

TABLE III. COMPARISON BETWEEN ACTUAL YIELD AND PREDICTED YIELD USING SVM

Measured Production	Forecast Production	Relative Error
0.001697	0.003527	1.078374
0.003307	0.005521	0.669489
0.001931	0.004945	1.560849
0.012992	0.011343	0.126924
0.011851	0.011132	0.060667
0.000834	0.005263	5.310552
0.007271	0.010004	0.375877
0.000834	0.000775	0.070743
0.000983	0.001306	0.328586
0.000826	0.000599	0.274818
0.005195	0.010992	1.115881
0.002943	0.003567	0.212029
0.022243	0.030932	0.390640
0.005987	0.005321	0.111241
0.089843	0.082113	0.086039
0.004124	0.003981	0.033220
0.001771	0.000755	0.573687

These three tables reveal that the error rate of the actual fishing yield measured against the forecast fishing yield is the lowest in Table 1 showing 18% using NRM. Table 2 shows 28% errors using RBF and Table 3 shows 30% errors using SVM. This implies that the proposed NRM method is more accurate in making fishing yield predictions. A further comparison is conducted involving five methods that have been used in fishing predictions with results in Table 4 and Table 5. This comparison shows NRM contains the lowest mean square error and the highest coefficiency. These comparison results support the conclusion that the NRM method surpasses all other existing methods in making fishing yield forecast.

TABLE IV. MEAN SQUARE ERROR OF 5 METHODS

RBF	SVM	SVM based on priori knowledge	Linear regression	NRM: new proposed method
0.043	0.037	0.035	0.058	0.033

TABLE V. COEFFICIENCY OF DETERMINATION OF 5 METHODS

RBF	SVM	SVM based on priori knowledge	Linear regression	NRM: new proposed method
0.4775	0.5513	0.5838	0.2889	0.6572

IV. CONCLUSION

This paper proposes a non-linear regression model (NRM) for fishing forecast. It employs cluster analysis and nonlinear regression to help forecast fishing yield based on marine environmental data. Measured data of Illex argentinus in Southwest Atlantic has been used for the experiment. The model has been verified for forecast accuracy and its result is compared with those applying the commonly used methods, RBF and SVM. The comparison confirms that NRM is more accurate in fishery prediction. There is speculation that if NRM is adopted it can lead to better efficiency and effectiveness in fishing.

ACKNOWLEDGMENTS

This work is supported by Innovation Program of Shanghai Municipal Education Commission (12ZZ162), and by Shanghai Municipal science and Technology Commission (12510502000), (14391901400). Their support is gratefully acknowledged.

REFERENCES

[1] Wei Fan, Xuesen Cui, Xinqiang Shen. Progress in fishing-ground analysis and fishing condition forecasting. Journal of Fisheries of China, 29(5), pp. 707-710, 1985.

[2] Zhengzhi Dong. The world oceans economic cephalopod biology. Jinan: Shandong Science and Technology press, pp. 131-134, 1991.

[3] Feng Chen, Xinjun Chen, Bilin Liu. Relationship between fishing ground of ommastrephes bartramii and vertical temperature structure in the northwestern Pacific Ocean. Journal of Shanghai Ocean University, 19(4), pp. 495-504, 2010.

[4] Xinjun Chen, Bilin Liu, Yuezhong Wang. Study on the distribution of production of Illex Argentinus and its relationship with sea surface temperature in the Southwest Atlantic Ocean in 2000. Journal of Zhanjiang Ocean University, 25(1), pp. 29-34, 2005.

[5] Feng Gao, Xinjun Chen, Jiangtao Fan. Implementation and verification of intelligent fishing ground forecasting of Illex Argentinus in the Southwest Atlantic. Journal of Shanghai Ocean University, 20(5), pp. 754-758, 2011.

[6] Rodhouse P G, Barton J, Hatfield EMC, Symon C. Illexargentinus: life cycle, population structure, and fisher. ICES Mar Sci Symp, 199(1), pp. 425-432, 1995.

[7] Bakun A, Csirke J. Environmental Processes and Recruitment Variability in Squid Recruitment Dynamics [M], Rome: FAO, pp.105-124, 1998.

[8] Weihua Song, Pengfei Xiong. Research on the developing of argentine shortfin squid-jig fishing in the Southwest Atlantic. Transactions of Oceanology and Limnology, (1), pp.62-68, 2002.

[9] Xinjun Chen. Fisheries Resources and Fisheries. Beijing: Ocean press, 2004.

[10] Global Capture Production. Food and Agriculture Organization of the United Nations. www.fao.org /fishery /statistics /global-capture-production /query /en.

[11] Yuexia Zhang, Zhongfeng Qiu, Yumei Wu. Predicting central fishing ground of scomber japonica in East China Sea based on case based reasoning. Marine Science, 33(6), pp. 8-11, 2009.