H-1723

ISC/05/MARLIN-WG/ 06

Preliminary analysis for area stratification and CPUE standardization of striped marlin caught by Japanese longline fishery in the north Pacific using tree regression models (TRM)

Hiroshi SHONO, Kotaro YOKAWA, Shelley CLARKE, Yukio TAKEUCHI,
Minoru KANAIWA* and Hirokazu SAITO

National Research Institute of Far Seas Fisheries
5-7-1, Orido, Shimizu, Shizuoka, Japan, 424-8633

* Tokyo University of Agriculture
196, Yasaka, Abashiri, Hokkaido, Japan, 099-2493

Abstract

Preliminary analyses were carried out using tree-regression model caught by Japanese longline fishery for striped marlin (5x5 degree square/monthly basis). Two kinds of analysis (area stratification and CPUE standardization) were performed and two types of algorithms for tree-regression models (CART and CHAID) were utilized in the calculation. Our results show that 1) Pattern of the partitioning for area stratification is largely dependent on the algorithm used. 2) Results of factorial experiment (extracted CPUE year trend and estimated gear effect) using simple method seem to be reasonable.

Key words: CPUE standardization, tree-regression models, striped marlin, area stratification, CHAID, CART, factorial experiment

Contents:

1. Introduction

Area stratification is an important issue on CPUE standardization for tuna species. However, it is generally difficult to divide the whole area into several sub-areas objectively and appropriately. Therefore, we examined the possibility to develop area stratifications (for CPUE analyses with GLM or habitat model and stock assessment through MULTIFAN-CL) based on CPUE of striped marlin caught by Japanese longline fishery in the north Pacific using tree-regression models (TRM). In addition, we carried out the CPUE standardization (i.e. estimated the CPUE year trend and gear effect) through factorial experiment based on the predicted values obtained from the TRM.

2. Data and model structure

We used the Japanese longline fishery data for striped marlin in the Pacific Ocean for 1975 to 2004 (5x5 degree square/monthly basis). Monthly CPUE (in number of catch per 1000 hooks) were calculated within 5x5 squares. The data from 1952 until 1974 was not utilized in our analyses because of the missing information on gear (i.e. hooks between floats, HBF).

We applied two typical algorithms of tree-regression model (TRM), CHAID and CART. The CHAID and CART analysis with TRM were performed through Clementine Tree package (Version 9.0) produced by SPSS Inc. The following five factors were incorporated as the explanatory variables. We set the above CPUE values for stripped marlin as a response variable.

Factors	Contents	Type of variable
Year	1975, 1976,...,2003, 2004	nominal categorical variable
Month	1,2,3,4,5,6,7,8,9,10,11,12	nominal categorical variable
Latitude	15S-0-35N	ordinal categorical variable
Longitude	110E-180-80W	ordinal categorical variable
Gear (HBF)	3-26	ordinal categorical variable

Remark) Latitude and Longitude were categorized in every 5 degrees.
(Of course, other Oceans were deleted from our calculations).

Criteria to continue further split into plural nodes with TRM are as follows:

- Number of data on parent nodes are more than or equal to 1000.
- Number of data on daughter nodes are more than or equal to 500.
- Depths of node (i.e. number of strata) are less than or equal to 5.

Chi-squared statistics and cost-complexity were also used as a criterion for partitioning in CHAID and CART, respectively.

3. Area stratification

We carried out area stratifications for CPUE standardization by GLM and/or habitat-model using two kinds of algorithm in tree-regression model (TRM), CHAID and CART. The basic concepts to partition the area considered are as follows.

- The areas that show similar CPUE values are grouped together.
- The areas on high CPUE are not partitioned into many parts.

(i.e. The boundary is set to on the areas in which CPUE shows lower value.)

We used only first several partitions (not fully split nodes) by TRM so as to get the appropriate (i.e. not so many) number of sub-areas for CPUE standardization.

The results of automatic area stratification based on CHAID and CART are shown in Figure 2 and 1-b, respectively. The split pattern of sub -areas between by CART and CHAID is quite different. The partitioning of the sub-area in the north of 25N and in the west of 120W by CART depends on month effect (Figure 1-b). Figure 1-a shows the first two partitioning by CART, which became rough area stratification for MLTIFAN-CL etc., the fully grown tree is shown in Figure 1-b.

Final area stratification based on CART and CHAID are quite different (Figure 1-b and 2). However, it seems to be rather difficult to find out this reason. In terms of statistical theory, CHAID algorithm is superior to CART because the problem of the multiple comparisons is still now unsolved. On the other hand, the split by CART algorithm is more adequate from the biological point of view (per. comm. Yokawa and Saito).

Therefore, we performed the two kinds of reliability check using the statistical method and compared the goodness of fit between CHAID and CART algorithm this time. One is a calculation of sum of square (SS) and another is the correlation analysis.

The SS is defined as the following formula.

$$SS = \sum_{y}\sum_{m}\sum_{L}\sum_{l}\sum_{g}\left(\hat{U}_{y,m,L,l,g} - U_{y,m,L,l,g}\right)^2 \qquad (0)$$

where *y*-year, m -month, *L*-latitude, *l*-longitude, *g*-gear(hooks between float) $U_{y,m,L,l}$ and $\hat{U}_{y,m,L,l}$ represent the observed and predicted CPUE, respectively.

Table 1 Sum of square and correlation coefficient based on CHAID and CART

Algorithm	SS (sum of square)	Correlation coefficient
CHAID	0.384	302,807
CART	0.307	321,869

We also computed the Pearson's correlation coefficient between observed CPUE value and the corresponding estimated one obtained from the TRM. It is found that the statistical performance on CHAID algorithm is a little better than that on CART and the difference seems not to be so large.

4. CPUE standardization

We calculated the standardized CPUE and gear (HBF) effects using the predicted CPUE (corresponding to each observed one) by tree regression models (TRM), CHAID and CART. Different from the GLM, it is generally difficult to do statistical factorial experiment such as extracting the CPUE year trend. Therefore, we used the simple method by Shono *et al.* (2001) for estimating of CPUE trend and gear effect using the fact that all explanatory variables are categorical ones. The idea of estimating year trend is as follow:

1. Calculating the predicted CPUE value $\hat{U}_{y,m,L,l,g}$ (for each year, month, 5x5 block and gear (hooks between floats (HBF)) corresponding to each observed one by TRM. Suffixes y, m and g represent year month and gear, respectively. Suffixes L and l indicate latitude and longitude that express 5 degrees.

2. Averaging the gear effects of predicted CPUE by following formula.

$$\hat{U}_{y,m,L,l} = \frac{1}{N_g}\sum_{g}\hat{U}_{y,m,L,l,g} \qquad (1)$$

for each set of suffix [*y, m, L, l*].

3. Averaging the month effects of predicted CPUE by following formula.

$$\hat{U}_{y,L,l} = \frac{1}{N_m}\sum_{m}\hat{U}_{y,m,L,l} \qquad (2)$$

for each set of suffix [*y, L, l*]. These averaging indices provide an annual CPUE for each 5x5 blocks.

4. Summing or averaging the area effects of predicted CPUE by following formula.

$$\hat{U}_{y} = \sum_{l}\sum_{L}\hat{U}_{y,L,l} \quad \text{Sum-type}$$

$$\text{or} \qquad (3)$$

$$\hat{U}_{y} = \frac{1}{N_l}\sum_{l}\left(\frac{1}{N_L}\sum_{L}\hat{U}_{y,L,l}\right) \quad \text{Ave-type}$$

These summing or averaging year-specific indices provide a CPUE trend.

Similar procedure was applied for estimating gear effects. Formulae (1)-(3) were replaced by the following (4)-(6).

$$\hat{U}_{y,L,l,g} = \frac{1}{N_m}\sum_{m}\hat{U}_{y,m,L,l,g} \qquad (4)$$

$$\hat{U}_{y,g} = \frac{1}{N_l}\sum_{l}\left(\frac{1}{N_L}\sum_{L}\hat{U}_{y,L,l,g}\right) \tag{5}$$

$$\hat{U}_{g} = \frac{1}{N_y}\sum_{y}\hat{U}_{y,g} \tag{6}$$

Thus, we obtain the gear-specific CPUE (i.e. CPUE in each number of HBF).

Figure 3 and 4 show the extracted CPUE year trends by CART and CHAID algorithm, respectively. Two CPUE series (RcpueSum and RcpueAve in Figures) in each figure shows the difference of the calculation way of area effects described in formula (3).

Figure 5 illustrate the gear effect on CPUE (i.e. estimated CPUE in each number of hooks between floats, HBP). These two extracted gear effects are rather similar and CPUE values by CHAID is more fluctuated that those by CART except for the part of small number in HBF.

The change of CPUE year trends by CHAID are more optimistic than that by CART and these results are a little similar to CPUE standardization based on GLM (Yokawa and Clarke, 2005) except for CPUE decrease for 1998 -1999 using CHAID. There overall CPUE year trend seems to be reasonable.

However, these CPUE values are not so robust and stable. In other words, predicted year trend of CPUE is dependent on the calculation method for categorical explanatory variables (e.g. latitude and longitude in this case). Therefore, more improvement of the factorial experiments including the case with continuous variable is necessary.

References

Breiman, L. J., Friedman, R. A., Olshen, R. A. and Stone, C., J. (1983). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley and Sons, New York.

Shono, H., Tsuji, S., Takahashi, N. and Itoh, T. (2001). Preliminary analysis for CPUE standardization and area stratification by tree-regression models. CCSBT-SC/0108/30. 17p.

Watters, G. and Deriso, R. (2000). Catch per unit of effort of bigeye tuna: a new analysis with regression tree and simulated annealing. Inter-American Tropical Tuna Commission Bulletin, Vol.21, No.8.

Yokawa, K. and Clarke, S. (2005). Standardizations of CPUE of striped marlin caught by Japanese offshore and distant water longliners in the north Pacific Submitted to the meeting of marlin working group for the ISC, November 15-21, ISC/05/MARLIN-WG, 9pp.

Appendix. Tree-regression model

A. Application of tree-regression models to the CPUE standardization

The analysis by generalized linear model (GLM) has been widely used to standardize CPUE. Several abundance indices of tuna species were calculated based on the results of GLM analysis with CPUE log-normal model. However, GLM method has several technical problems as follow:

- Missing (or zero) data make an estimation of interactions difficult using GLM. Spatial and temporal distribution pattern for many tuna species changes drastically from year to year and area to area. Therefore, estimation of interactions like (Year)*(Area) is difficult in most cases by GLM unless integrate the categories about year and area widely.
- Due to the constant term added to response variable, CPUE, confidence interval has a bias, in the CPUE model with log-normal error structure.

In contrast to GLM, Tree-regression models (TRM) automatically detect and extract important factors to separate data into groups showing similar pattern. The advantages of TRM are as follow:

- No specific assumption it need for data distribution pattern, such as constant seasonal and timely pattern, and statistical structure etc.
- We can avoid the problems of interactions and constant term, which are related to missing data or zero catch.
- Interpretation of CPUE (and abundance index) with TRM is easier than that by GLM. For instance, to extract the effect of the lower or terminal node usually means to check more than two-way interactions in the GLM analysis.

TRM has been often used as a method for data mining, recently. In the filed of fish population dynamics, Watters and Deriso (2000) applied the algorithm of TRM to the CPUE analysis for bigeye tuna in the Pacific Ocean.

B. Typical algorithm of tree-regression models, CHAID and CART

CHAID algorithm (Hartigan, 1975) is one of the typical algorithms of tree-regression model (TRM) partitioning each node based on the values of chi-squared statistics. The characteristics of CHAID algorithm are as follow:

- A split into more than two nodes is allowed on each turning point.
- Continuous variables are transformed into categorical one.

CART algorithm (Breiman *et al.*, 1984) is one of well-known algorithm that split the data into two nodes so as to maximize the difference of mean value of predicted variable. In CART model, pruning is usually conducted based on the criterion of cost-complexity after the tree is fully grown up and the problem of statistical multiple comparisons are not solved.

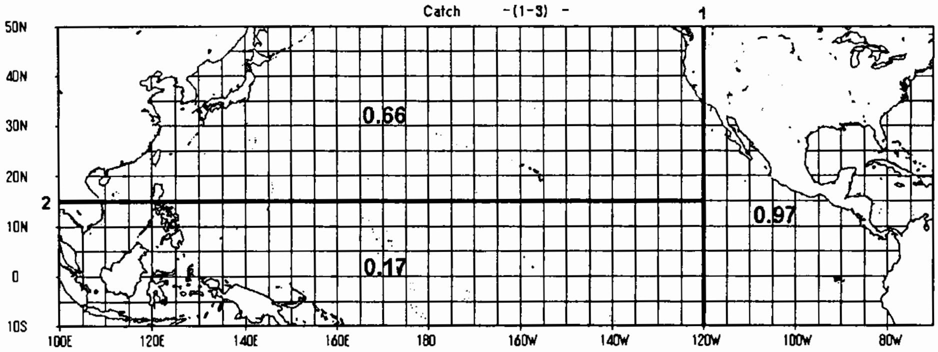

Figure 1-a. Area stratification (for MULTIFAN-CL etc.) by CART algorithm.

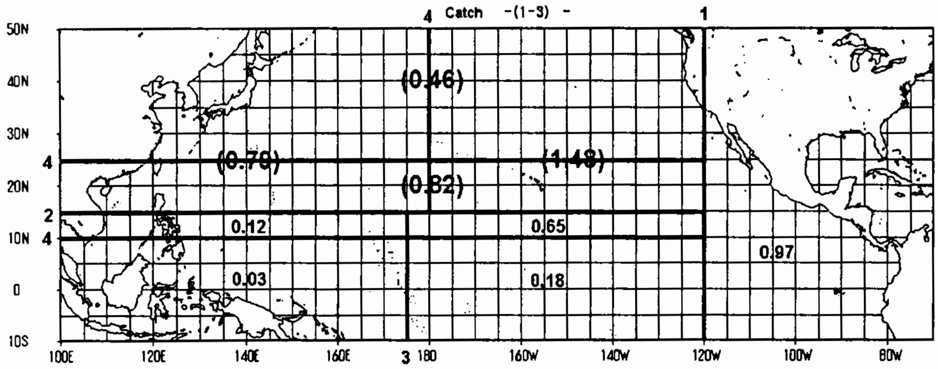

Figure 1-b. Area stratification (for GLM / habitat-model) by CART algorithm.

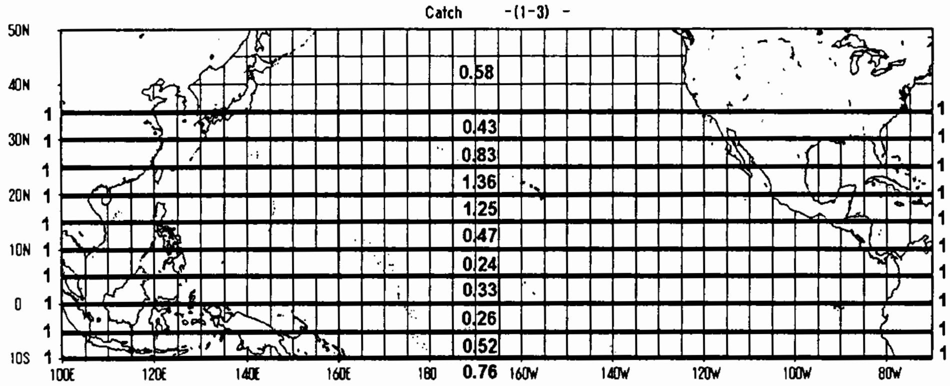

Figure 2. Area stratification by CHAID algorithm.

Remark) The numbers out of the map shows the order of partitioning using the TRM. The numbers in the map shows the estimated CPUE values using the TRM. The numbers of parenthetic in Figure 1-b shows that these CPUE values depend on the month effect as the following tables:

Month-4, 5, 6, 10	110E≦Longitude＜180	180≧Longitude＞120W
25N≦Latitude＜40N	0.79	1.48
15N≦Latitude＜25N	0.79	1.48

Month-1,2,3,7,8,9,11,12	110E≦Longitude＜180	180≧Longitude＞120W
25N≦Latitude＜40N	0.46	0.46
15N≦Latitude＜25N	0.82	0.82

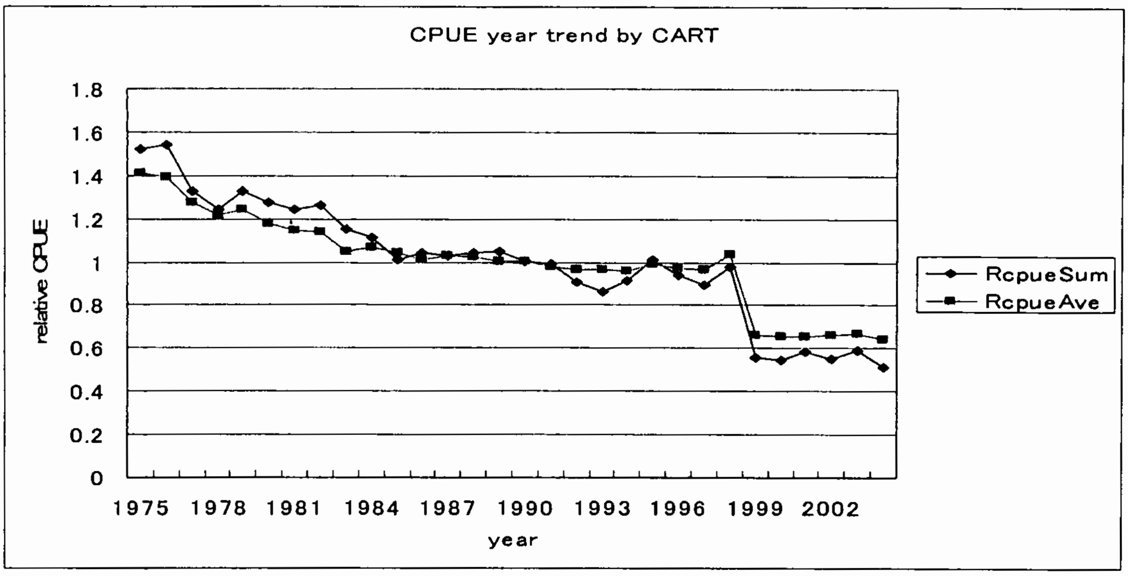

Figure 3 CPUE year trends by CART algorithm (Average value is set to 1).

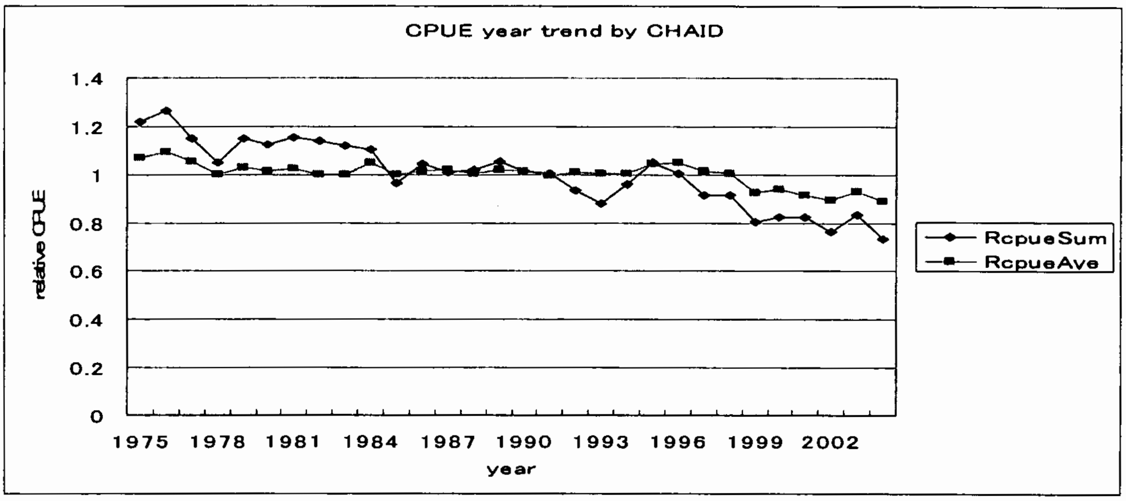

Figure 4 CPUE year trends by CHAID algorithm (Average value is set to 1).

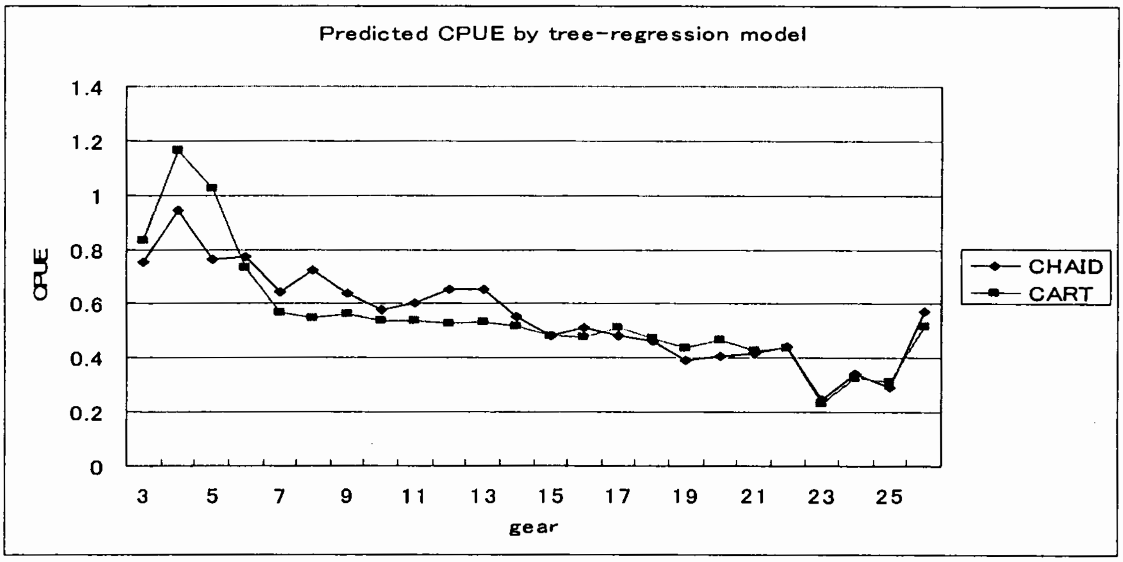

Figure 5 Estimated CPUE (by CHAID&CART) in each number of NBF (gear)