CCSBT-ESC/0409/43

Attempt for multiple imputation of SBT-CPUE using new statistical method

新しい統計手法を使用したミナミマグロ CPUE の多数回補完の試み

Hiroshi SHONO（庄野 宏）

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

Abstract: Three types of CPUE analyses were carried out by multiple imputations method using the Japanese longline fisheries data for southern bluefin tuna (5x5 degree square/monthly basis). The estimation of the missing CPUE in the core-area where many fishery data are available using the propensity score method were performed and compared with EFP data. Although the fitting is not so good, EFP data do not become the adequate supervised one. We performed the reliability check of the interpolated CPUE by "n-fold cross-validation" using the same data as those used in the CPUE standardization by GLM. The accuracy seems not to be so good because of some extreme outliers. Preliminary statistical analyses with a view to predict CPUE in non-fished were done by the propensity score method. As a result, it was found that the year trend of abundance index by multiple imputation method is rather different from those in CS and VS-type by GLM.

要旨：日本のミナミマグロはえ縄漁業の 5 度区画・月別に集約したデータを使用して多数回補完法による 3 種類の CPUE 解析が行われた。多くの漁業データが利用可能なコア・エリアにおける欠損 CPUE の推定が傾向スコア法を用いて行われ、(推定された CPUE が)EFP データと比較された。そのフィッティングはそれほど良くないが、EFP データは必ずしも適切な教師付きデータになっていない。一般化線形モデルによる CPUE 標準化と同じデータを用いて、“ n-fold クロス・バリデーション” による補間された CPUE の信頼性の検証が行われた。その精度は幾つかの極端な外れ値のために、さほど良くはなっていない。漁獲のないエリアの CPUE 予測を目的とした、傾向スコア法による予備的な統計解析が行われた。その結果、多数回補完法による資源量指数の年トレンドが CS 及び VS タイプの一般化線形モデルによるそれらとかなり異なることが判明した。

1. Introduction

The difference of interpretation about SBT CPUE in the non-fished area between constant square (CS) and variable square (VS) have greatly affected not only the relative year trends of abundance index but also the result of stock assessment in which these CPUE indices are utilized as an input data. Therefore, five abundance indices including two types of the weighted mean of CS and VS (W0.5 and W0.8) are defined as the input for the management procedure.

Several statistical methods for CPUE interpolation in the non-fished area have been suggested so far in CCSBT. Australian scientists tried to calculate the so-called "geo-statistical CPUE" using the method of repeated measure, which is a kind of mixed model (Toscas and Thomas, 1998). We predicted the CPUE in 5x5 degree blocks in which there is no observed data by the approaches for data-mining such as tree regression models and neural networks (Shono *et al.*, 2001: Shono, 2002). However, the discussion in the meeting and the reliability check by cross-validation about these methods for CPUE interpolation are not sufficiently done.

In this paper, we carried out the three types of CPUE analyses using several multiple imputation methods. We mainly utilized the called propensity score method and performed the reliability check of interpolated CPUE obtained from these new statistical methods through cross-validation or comparison with the testing data for verification. These are as follows:

1) CPUE prediction in the "core-area" (including "ST-window" series) by propensity score method and cross-check with the corresponding EPF data.
2) Reliability check of interpolated CPUE by n-fold cross-validation using the same data as used in the GLM analyses. CPUE estimation of the missing parts was done by the multiple imputation method using the combination of MCMC and EM-algorithm suggested by Little and Rubin (2002).
3) Multiple CPUE standardizations by GLM (i.e. extraction of the year trends of CPUE) using the completely interpolated CPUE data by the propensity score method.

Propensity score method (Rosenbaun and Rubin, 1983: 1985) is a statistical way for imputation: The procedure is as follow:

1) Express the probability that the data is observed using a statistical model such as logit-model.
2) Divide the whole data into several sub-data based on the magnitude of "propensity score"
3) Interpolate the missing data using so-called "Aproximate Baysian bootstrap" in each sub-data.

Because the interpolated values obtained from this method have some uncertainty, multiple imputations are generally performed. In this paper, we carried out the CPUE imputations for five times.

2. CPUE imputation in the core-area

At first, we applied the calculation of multiple imputations using propensity score method to SBT CPUE for age4+ (age 4 and older) of non-fished area in the following "core area" where fishery data can be securely available even when the pattern of SBT fishery changes.

Year: 1993-2003
Month: 4, 5, 6, 7, 8, 9
Statistical area: 8 and 9
Remark): We used the blocks in above areas categorized in every 5 degrees.

These cells include the "ST-windows" and some of them are corresponding to spatiotemporal window in which EFP experiments have been carried out in the last 1990s. Therefore, we verified the interpolated CPUE values using the corresponding EFP data.

Figure 1 shows the mean values of predicted CPUE for five times imputations and corresponding (i.e. in the same spatiotemporal window) EFP data. CPUE values by multiple imputations are distributed around two or three regardless of the EFP data and the precision is not so good. However, CPUE obtained from the EFP experiment do not fit the corresponding (i.e. in the same spatiotemporal window) logbook data (Figure 2). Therefore, it is found that CPUE by EFP is not necessary appropriate as supervised data.

3. Reliability checks by n-fold cross-validation

Next, we checked the reliability of multiple imputations by so-called “n-fold cross-validation” using the Japanese longline fishery data for SBT from 1969 to 2003 (5x5 degree square/monthly basis) in all fished area, which is exactly the same as those used in the calculation of abundance indices for the 2004 SAG/SC assessment by generalized linear models (Tsuji *et al.*, 2004). Monthly CPUE (in number of catch for age 4+ (age 4 and older) per 1000 hooks were utilized within 5x5 squares.

In this paper, we adopted the 5-fold cross-validation. The procedure of the calculation is as follows:

1) Divide the observed CPUE (7,697 records) into five sub datasets randomly and assigned a number of these sub-sets in turn.
2) Choose a sub-set and regarded CPUE as the missing values in the dataset.
3) Do multiple imputations of the missing CPUE and check the precision using Pearson’s correlation coefficient and the following Index (equation (1)) like chi-squared statistics (or relative mean squared error).
   Remark: We used the combination of MCMC and EM-algorithm (Little and Rubin, 2002) instead of propensity score method in the calculation.
4) Replicate step 2) and 3) to all sub datasets (i.e. replication for five times).

$$Index = \sum_{i=1}^{n} \frac{(\bar{X}_i - Y_i)^2}{(\bar{X}_i + Y_i)/2} \qquad (1)$$

where

$\bar{X}_i$: simple mean of predicted CPUE, Y_i: supervised data corresponding to $\bar{X}_i$

Figure 3 shows the overall relationship between simple mean of interpolated CPUE and corresponding supervised data (i.e. observed CPUE). Table 1 shows the Pearson’s correlation coefficient and the value of Index defined by equation (1) in each random sub-dataset. Frequency (histogram) and cumulative distribution of this Index were illustrated in Figure 4.

Judging from these values and figures, the reliability of the predicted

CPUE obtained from the multiple imputations seems not to be so good. However these values of Pearson's R and the Index in Table 1 come under the strong influence of the small number of extreme "outliers". Therefore, the precision of these estimates can be improved by removing some outliers.

We also calculated the CPUE year trends by the following formula (2) replacing the (shadow) missing data with the interpolated CPUE in each sub dataset (Figure 5). In the figure, "original" and "predicted" shows the CPUE trends obtained from the all observed CPUE and all estimated one, respectively.

$$CPUE_{year} = \frac{1}{N_m}\left[\sum_{month=1}^{N_m}\frac{1}{N_l}\left\{\sum_{longitude=1}^{N_l}\frac{1}{N_L}\left(\sum_{Latitude=1}^{N_L}CPUE_{year,month,longitude,Laitude}\right)\right\}\right] \qquad (2)$$

When the ratio of the testing data (i.e. missing values) is low such as "CPUE(Ⅰ)—CPUE(Ⅴ): (where the ratio of missing is 20 percent)" in Figure 5, these CPUE trends are similar to the "original" one. On the other hand, the trend of "predicted" used only the predicted values are quite different from (and flatter than) that of "original" (Figure 5). Therefore, it seems to be necessary to do more discussion and study in detail regarding the proportion of "training-data" and "testing-data" (i.e. data for validation) as well as the method of "cross-validation" in such kind of multiple imputations.

4. Integrated results of our CPUE analyses

At last, we tried to predict the CPUE of non-fished area (i.e. 5x5 blocks in which there is no observed data: 22,123 records) using the propensity score method based on the same supervised CPUE data (7,697 records) as used in the previous Section 3 (n-fold cross-validation) of this paper. We used the 5x5 blocks there exists some fishing effort in the past as the non-fished area, which is the same as used for the calculation of the abundance index in constant square (CS) type.

We carried out the multiple imputations for five times in this paper. Simple GLM analyses defined by the following equation (2) were independently performed on the basis of the above interpolated CPUE data.

After that, we integrate the results of our calculations for five times (based on the five dataset in which the missing CPUE data were fully interpolated).

Figure 6 shows the year trends of standardized CPUE using the completed five datasets interpolated by the multiple imputation methods. These trends are very similar although each predicted values of our experiments for five times are rather different. However, an integrated year trend of abundance index, which is a simple mean of five standardized CPUE by GLM and where average of CPUE values is set to 1.0, is flatter than those in constant square (CS) and variable square (VS) type (Figure 7). The difference of these CPUE trends is rather large.

Acknowledgement

We acknowledge Dr. K. Hiramatsu, National Research Institute of Far Seas Fisheries, for his useful technical advice.

References

Anonymous. (1998). Report of the first meeting of the scientific assessment group (SAG), 41 pp.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data* (second edition). John Wiley & Sons, New York. 381pp.

Rosenbaun, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effect. *Biomatrika* 70, p.41-55.

Rosenbaun, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score. *Ann. Statist.* 39, p.33-38.

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, 17 pp.

Shono, H. (2002). Attempts for estimation of standardized CPUE by tree-regression modes and neural networks. CCSBT-ESC/0209/38. 18 pp.

Toscas, P. and Thomas, M. (1998). Spatial analysis of southern bluefin tuna catch per unit effort data: A best linear unbiased predictor approach. CCSBT-SC/9807/10, 33 pp.

Tsuji, S., Itoh, T. and Takahashi, N. (2004). Summary of fisheries indicators in 2004. CCSBT-ESC/0409/34, 15pp.

Appendix. Table and Figures.

Table 1. Correlation coefficient and above Index corresponding each sub-set.

Sub-Set	I	II	III	IV	V
Pearson's	0.287600	0.313245	0.320535	0.252071	0.280838
Index	3.837353	3.916269	3.668749	3.679176	3.754673

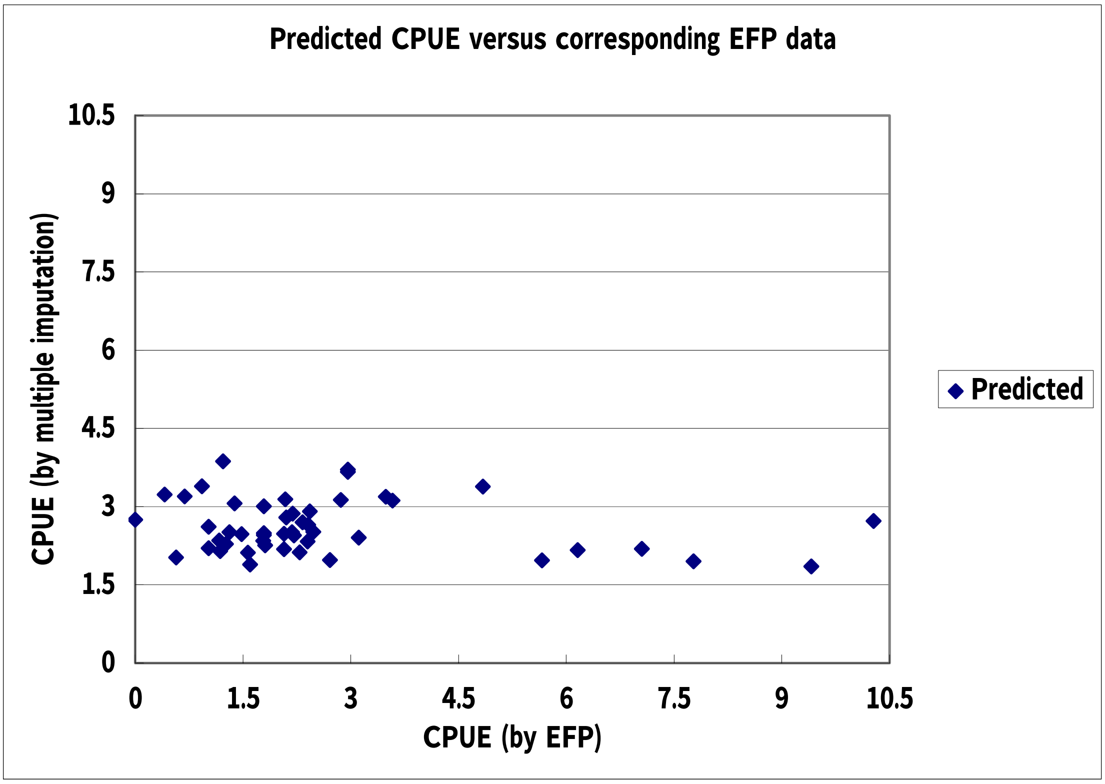

Figure 1. Average CPUE (based on propensity score) vs. CPUE (by EFP).

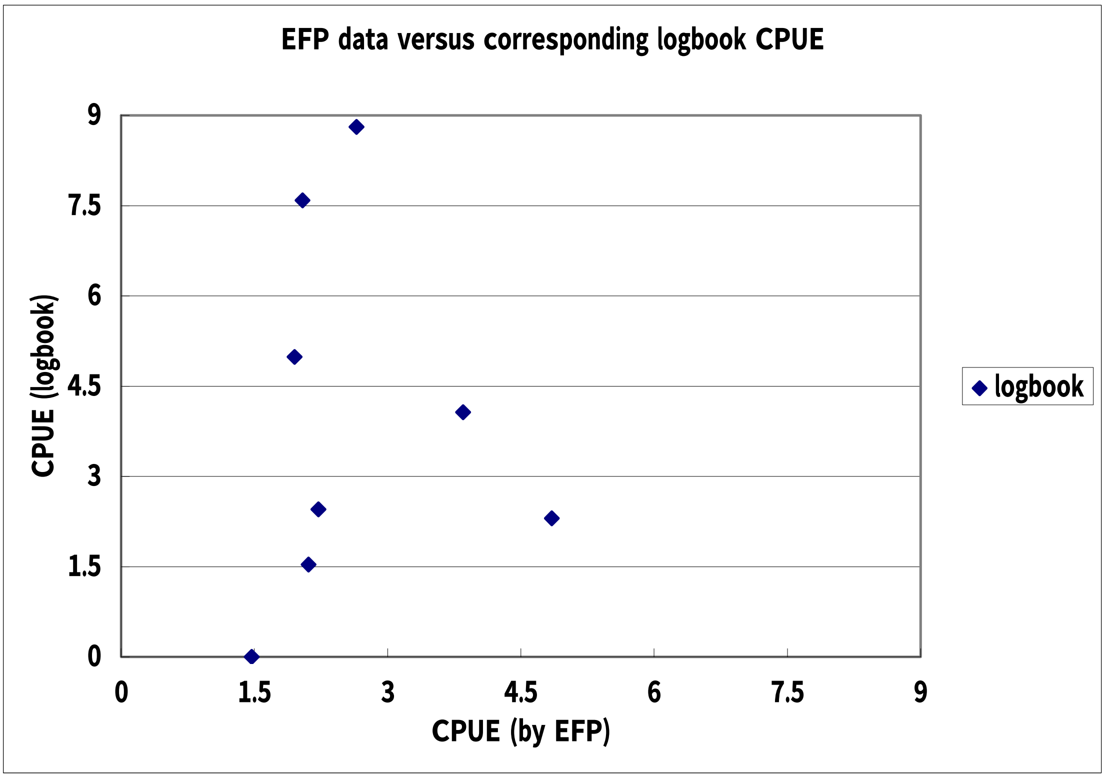

Figure 2. CPUE (by EFP) versus the corresponding CPUE (by logbook data).

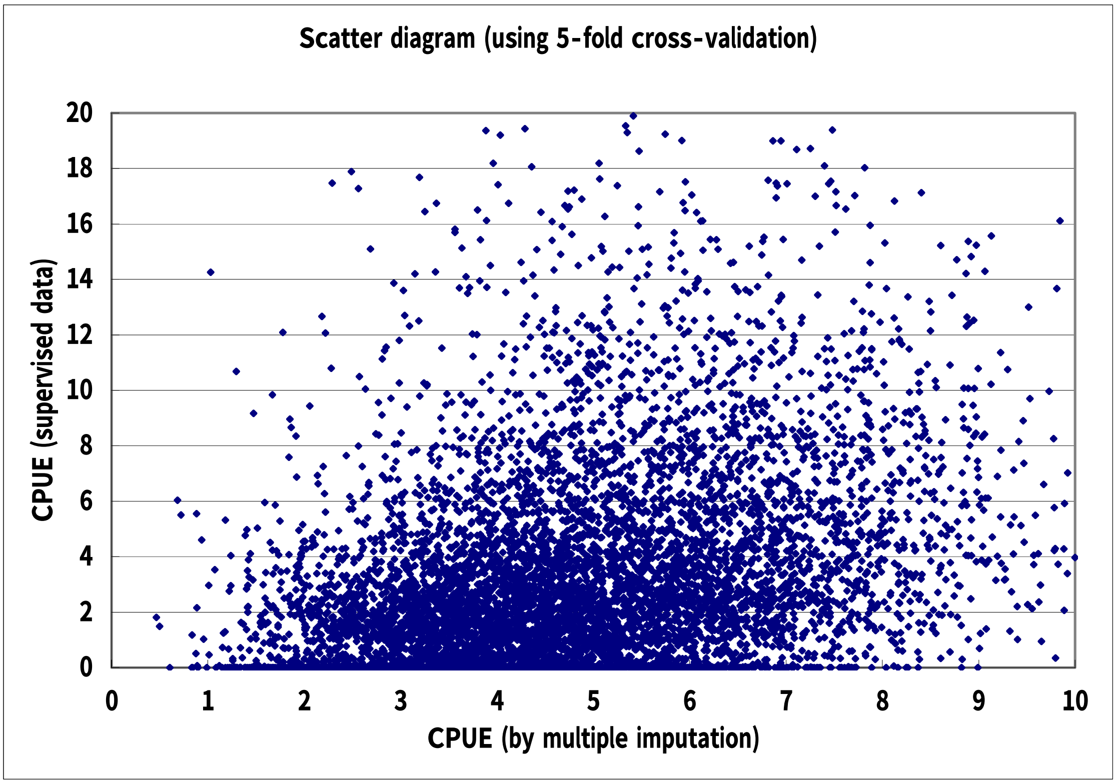

Figure 3. Average of predicted CPUE versus corresponding supervised data.

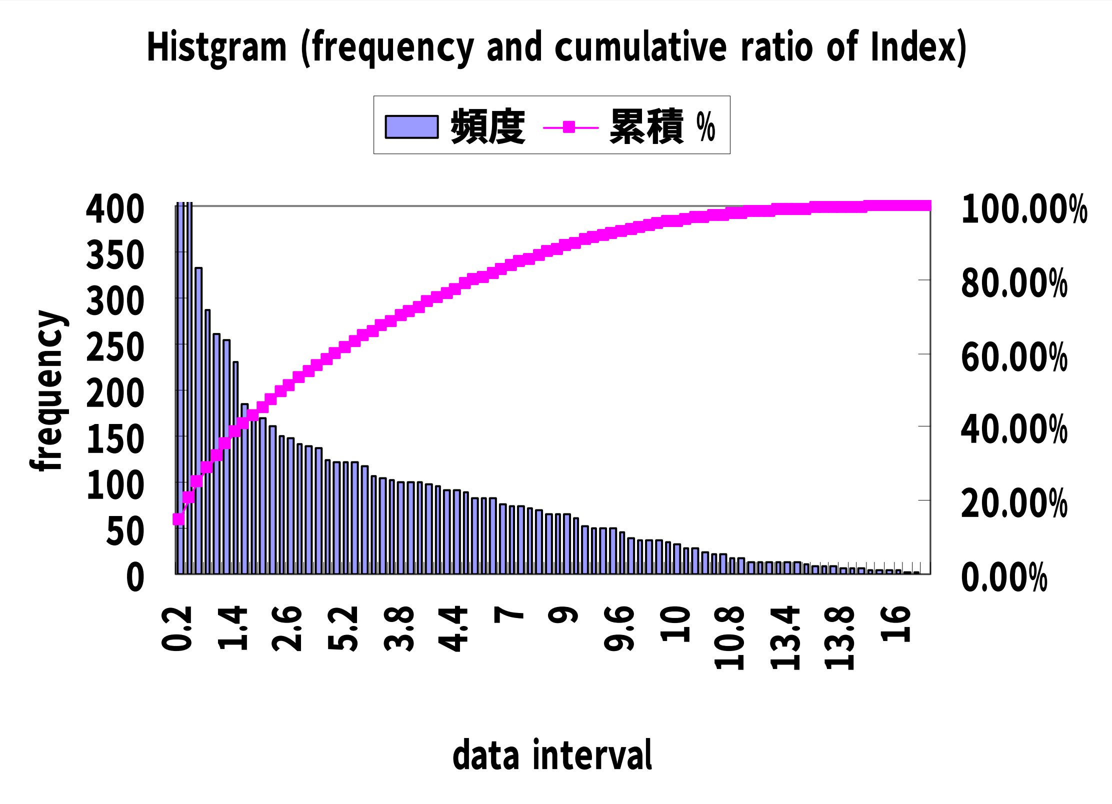

Figure 4. Frequency and cumulative ratio of the histgram of the Index value.

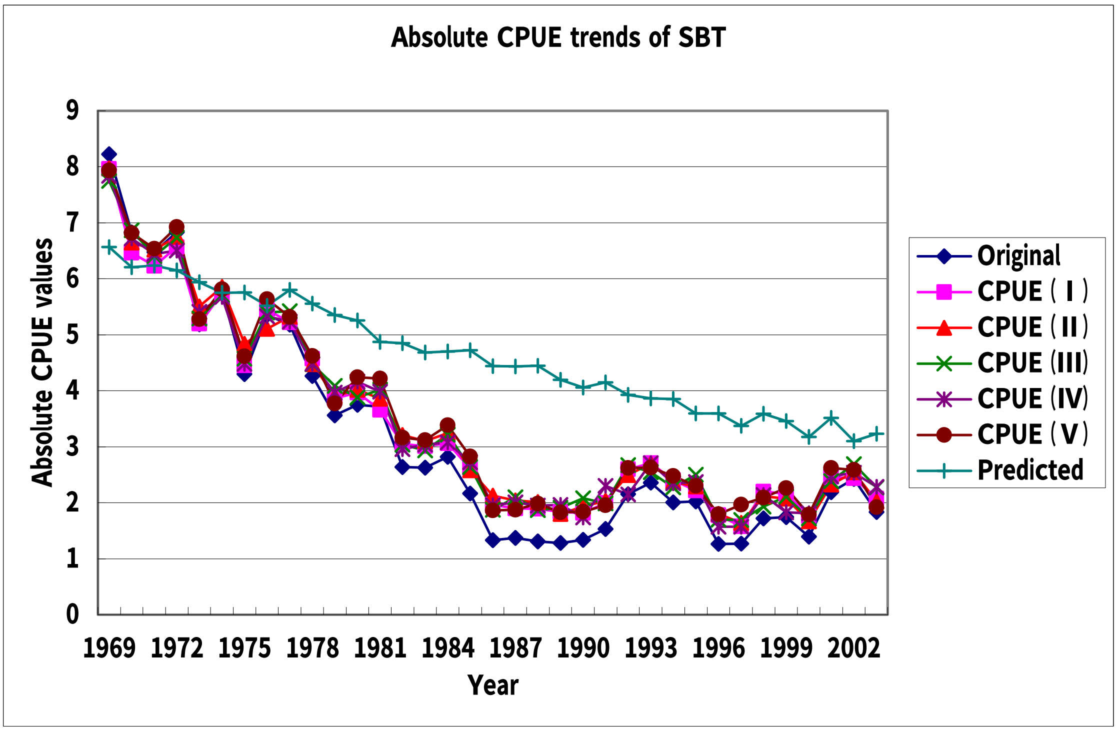

Figure 5. Year trends of absolute CPUE interpolated in each sub-set (Ⅰ－Ⅴ).

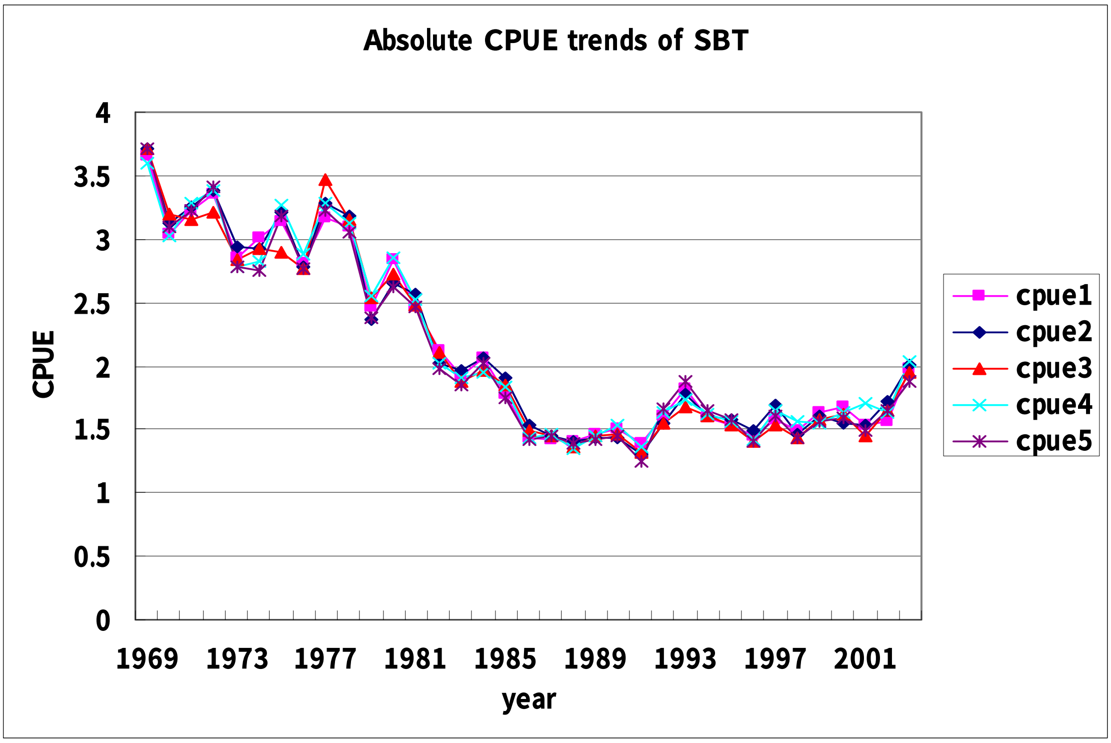

Figure 6. Year trends of absolute CPUE by GLM using interpolated dataset.

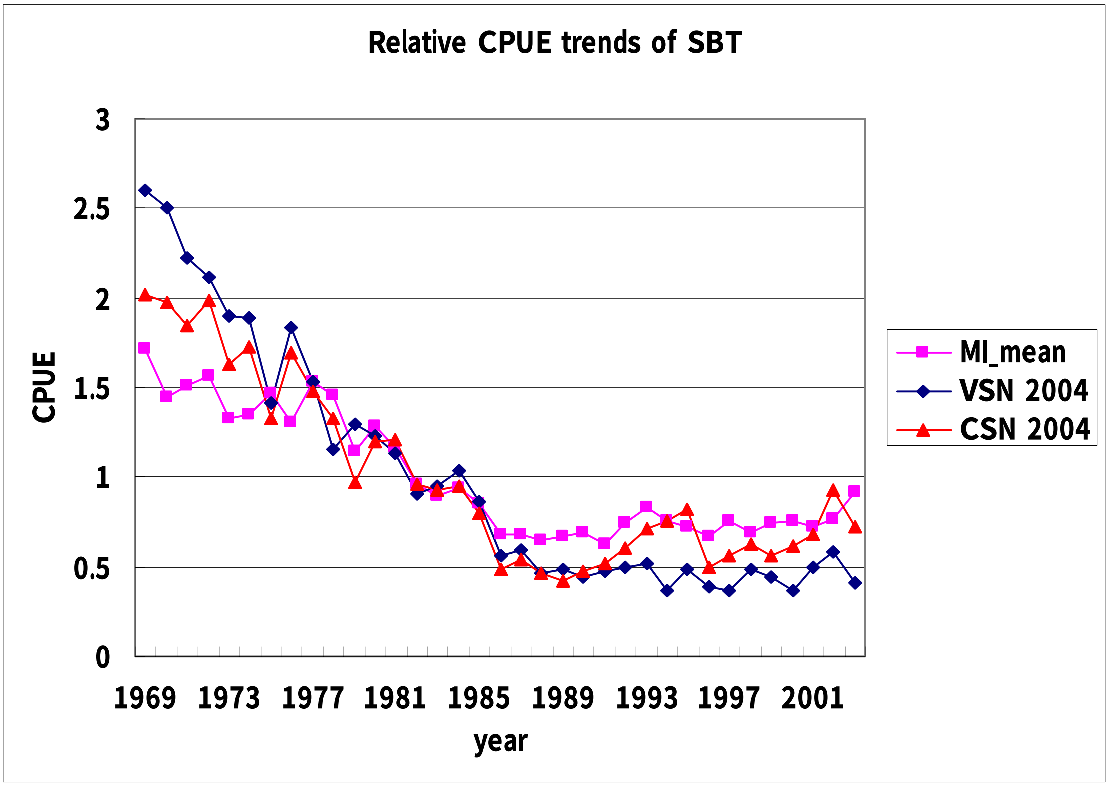

Figure 7. Year trends of relative CPUE by GLM (Compared with CS and VS).