Data filtering of Japanese logbook data in the Indian Ocean for analysis of species-specific shark's data from 1993 to 2018

Mikihiko Kai[1]

[1]National Research Institute of Far Seas Fisheries,
Japan Fisheries Research and Education Agency
5-7-1 Orido, Shimizu-ku, Shizuoka 424-8633, JAPAN

Key words

Japanese longline fishery, Data-filtering, Reporting ratio, Sharks

Abstract

Japanese logbook data have high spatial and temporal coverages in the Indian Ocean compare to those of observer data. However, the logbook data includes a large number of under-reporting catches for sharks, which makes it difficult to use directly the logbook data for the estimation of annual catch rates for sharks in the Indian Ocean. In order to solve the issue, the author used a statistical data method to filter Japanese logbook data and removed the systematic annual trends by reducing the set-by-set data with low reporting rates of catch for sharks using information on observer data. The reliability of the filtering method was validated using annual nominal CPUEs of tunas and sharks.

Introduction

Japanese longline fisheries have been operating in the Indian Ocean since 1950s to catch tunas and tuna-like species. However, there is no species-specific data of sharks until 1990s. National Research Institute of Far Seas Fisheries (NRIFSF) had commenced to collect information about the species of sharks since 1992. However, the logbook data have an issue on the low reporting rates of sharks due to the nature of bycatch species. Meanwhile, the observer data has no issue on the reporting rates, but the coverage of observer data is much lower than that of logbook data (less than 8 %).

Catch Per Unit Effort (CPUE) of sharks is commonly used as the abundance indices of sharks (**Kai *et al.*, 2017a,b**). The higher the temporal and spatial coverages of data are, the more reliable the CPUEs are. In light of the spatial and temporal coverage of data, the use of logbook data is more preferable than observer data to estimate the annual CPUEs with high precision, if the set-by-set logbook data of low-reporting rates are removed. Recent developed filtering methods for set-by-set logbook data (**Hoyle *et al.*, 2017; Kai, 2019**) are useful ways to remove set-by-set logbook data with under-reporting of actual shark catches. The main purpose of this working document paper is to apply a data filtering method to Japanese logbook data collected in Indian Ocean by NRIFSF from 1993 to 2018 to filter set-by-set logbook data with low reporting rates for CPUE standardization of sharks in the Indian Ocean.

Material and Method

1) Data sources

Japanese longline logbook data in the Indian Ocean from 1993 to 2018 and its observer data were used to filter out the data with under-reporting. The both data in 1992 were not used in this study due to extremely low number of set-by-set logbook data. Both data include catch data (catch number of tuna and tuna like species, and species-specific sharks), effort data (number of hooks), spatial and temporal information (latitude, longitude, year, and month), gear configuration (number of hooks between floats; HBF), and vessel name, etc.. Since the information about vessel trip was not included in the logbook data, the vessel trip was arbitrarily assigned using the combinations of operational year, month and vessel name. The data of species-specific sharks include blue shark, shortfin mako, porbeagle shark, thresher sharks, oceanic whitetip-shark and other sharks.

2) Data filtering

Incomplete and insufficient data were filtered, as were sets that have little or no information about HBF and locations (latitude and longitude), numbers of hooks that were less than 800 and more than 4200, HBF that were less than 4 and more than 40, and operations that were conducted in waters other than the Indian Ocean (50°S –20 °N and 20–145 °E except for the northeastern waters near Southeast Asian; **Fig. A1**). In this paper, this filtering step is referred to as "preliminary filtering".

Set-by-set logbook data with under-reporting were further removed using the following filtering method. **Hoyle et al. (2017)** developed a filtering method based on previous studies (**Nakano and Clarke, 2006; Walsh and Kleiber, 2001; Walsh *et al.*, 2002**) and defined three terms:

I. $RI \text{ (Reported incidence)} = \frac{Number\ of\ sets\ with\ sharks\ recorded}{Total\ number\ of\ sets}$,

II. $SP \text{ (Probability of shark presence in the catch)} = \frac{Number\ of\ sets\ with\ sharks\ caught}{Total\ number\ of\ sets}$,

III. $SR \text{ (Shark reporting reliability)} = \frac{Number\ of\ sets\ with\ sharks\ recorded}{Number\ of\ sets\ with\ sharks\ caught}$,

where *RI* is equivalent to reporting rates per cruise, or vessel trip (**Nakano and Clarke, 2006**). The expected value of *RI* is equal to *SP* × *SR*. The author assumed that *SR* would be equal to 1 (i.e. *RI* = *SP*) for observer data from 1993 to 2018.

The procedure for this data filtering is as follows:

(i) *RI* (RI_{LB} and RI_{OB}) was calculated using logbook data (LB) and observer data (OB) for each vessel trip, respectively;

(ii) Given that RI_{OB} (RI_{true}) is equivalent to SP_{OB} (SP_{true}) and these values were estimated using a generalized linear model (GLM) with binomial error distribution:

$$SP_{\mathrm{true}} = RI_{\mathrm{true}} \sim year + month + lat15 + lon15 + hbf \qquad (1)$$

where *year* and *month* are temporal effect, *lat*15 and *lon*15 are spatial effects denoting latitude and longitude on a 15-degree grid, respectively, but the latitude were finally gather into three (30-50°S, 16-30°S, 15°S-25°N) to cover a lacking of data and *hbf* is gear effect denoting hooks between floats (<9, 9-14, 15-19, 19<). These explanatory variables were arbitrarily given in consideration of the balance of each effect;

(iii) The best model was selected using a stepwise AIC.

(iv) The estimates of the coefficient for the best model were used to predict the expected values of *SP* for each set-by-set logbook data (SP_{LB});

(v) SR_{LB} by vessel trip was calculated by averaging the set-by-set data of respective RI_{LB} and SP_{LB} by vessel trip;

(vi) To validate the accuracy of the prediction given by the model and to determine the threshold, 80% confidence intervals of SR_{OB} (SR_{true}) were estimated using the output (i.e. coefficients) from Eq. (1);

(vii) The data for SR_{LB} was cut off if it was smaller than the lower bound of the 80% confidence intervals of SR_{OB}.

In this paper, this filtering step is referred to as “follow-up filtering”. For Eq. (1), vessel effect was not modeled because there was little vessel overlap during the periods. Positive catch ratios of sharks per cruise against the explanatory variables are shown in **Fig. 1**.

3) Validation of estimates

To validate the reliability of the filtering method, annual nominal CPUE of tunas (Southern Bluefin tuna, Bigeye tuna, Yellowfin tuna, Albacore) were compared using the logbook data with and without filtering. In addition, annual nominal CPUE of sharks were also compared to show the different effect of the filtering on the annual nominal CPUE between tunas and sharks.

Results

The preliminary filtering for logbook data and observer data reduced the number of records for this analysis from 603,427 sets to 595,784 sets, and from 14,412 sets to 13,764 sets, respectively. The follow-up filtering for logbook data reduced the number of records for this analysis from 595,784 sets representing 27,795 trips to 95,914 sets representing 4,696 trips.

The full model (M-6) was selected as the most parsimonious model in the analyses of the follow-up filtering (**Table 1**), and all factors of the deviance table were statistically significant

(**Table 2**). The lower and upper 80% confidence intervals of SR_{LB} were estimated as 0.782 and 1.151, respectively (**Fig. 2**), and lower bound was used as a cut-off point as shown in **Kai (2019)**. The threshold (i.e. 0.782) appeared to be reasonable because the reporting rates of catch for sharks increased and were close to 1 (**Fig.3c**). Annual nominal CPUE of shortfin mako (SFM) and blue shark (BSH) for logbook data was largely changed after filtering, respectively (**Figs.4c,f**) due to the reductions of annual catch as well as annual fishing effort (i.e., number of hooks) (**Fig. 4a,b,d,and e**). Annual nominal CPUE of SFM and BSH for observer data showed a large different trends compared to those for logbook data after filtering (**Figs.4c,f**). Although the annual nominal CPUEs for tunas showed a similar trend between two data-set with and without filtering, the annual nominal CPUEs for sharks were significantly different between them (**Fig. 5**). These results suggested that the filtering method used in this study is reasonable because there was a small impact of data filtering on the annual trend of tunas even if a large number of data-sets were removed from the analysis.

Discussions

In this study, a statistical method (**Hoyle *et al.*, 2017**) was used to filter Japanese logbook data from 1993 to 2018 in the Indian Ocean. One of the biggest issues of Japanese logbook data in the Indian Ocean was the low reporting rates of catch for sharks especially for the years before 2008 (**Fig. 1e**; **Fig. 3b**). However, the filtering method completely removed the systematic annual trends by reducing the set-by-set data with low reporting rates of catch for sharks (**Fig. 3c**) using information about observer data (**Fig. 3a**). In addition, the reliability of the filtering method was validated using annual nominal CPUEs of tunas and sharks (**Fig. 5**).

It was considered that the differences of annual nominal CPUEs for SFM and BSH between observer data and logbook data after filtering (**Figs. 4c,f**) were caused by the differences of the spatial coverages of the data-sets (**Figs. A1a,b**). The number of records and spatial coverages for logbook data overwhelmed those for observer data (**Figs. A1**). It is therefore desirable for the CPUE standardization of pelagic sharks in the Indian Ocean to use the logbook data after filtering.

To remove the logbook data with low reporting rates, lower bound of 80% confidence interval was used as a cut-off point in consideration with the number of records (to maintain higher spatiotemporal coverages) and a value of shark reporting reliability (to maintain a value closer to 1). The value of threshold has a large impact on the reporting rates of catch for sharks (**Fig. A2**). Lower value of threshold smaller than 80% confidence interval increases the reporting rates of catch for sharks, while higher value of threshold larger than 80% confidence interval decreases the reporting rates of catch for sharks. Three annual CPUEs were compared to evaluate the effect of the different cut off points. The trends in the annual nominal CPUE were almost similar among them (**Fig. A3**). These results suggested that the value of threshold is a small effect on the trends in the annual nominal CPUE.

In the previous study, **Hoyle *et al.* (2017)** used a cluster as a categorical variable in the binomial GLM to consider the targeting effects. In the Indian Ocean, Japanese vessels targets Southern bluefin tuna in the temperate waters and Bigeye tuna and Yellowfin tuna in the tropical waters. The target shift may change the catchability of vessels for sharks. Therefore, it is essential to consider the targeting effect in the analysis of the fishery-dependent data such as Japanese logbook data. However, the target shift during the fishing operation by Japanese longliner in Indian Ocean is uncommon. Rather, there is a clear difference for the target species between temperate and tropical areas as mentioned above. It is therefore enough to have considered the effect of spatial and temporal distributions as well as hooks between floats in Eq (1).

References

Hoyle, S.D., Semba, Y., Kai, M., Okamoto, H., 2017. Development of Southern Hemisphere Porbeagle Shark Stock Abundance Indicators Using Japanese Commercial and Survey Data. New Zealand Fisheries Assessment Report 2017/07.

Kai, M., Thorson, J.T., Piner, K.R., Maunder, M.N., 2017a. Predicting the spatio-temporal distributions of pelagic sharks in the western and central North Pacific. Fish. Oceanogra. 26, 569–582. https://doi.org/10.1111/fog-12217.

Kai, M., Thorson, J.T., Piner, K.R., Maunder, M.N., 2017b. Spatio-temporal variation in size-structured populations using fishery data: an application to shortfin mako (*Isurus oxyrinchus*) in the Pacific Ocean. Can. J. Fish. Aquat. Sci. 74, 1765–1780. https://doi.org/10.1139/cjfas-2016-0327.

Kai, M. 2019. Spatio-temporal changes in catch rates of pelagic sharks caught by Japanese research and training vessels in the western and central North Pacific. Fish. Res. 216, 177–195.

Nakano, H., Clarke, S., 2006. Filtering method for obtaining stock indices by shark species from species-combined logbook data in tuna longline fisheries. Fish. Sci. 72, 322–332.

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. Fish. Res. 53, 115–131. https://doi.org/10.1016/S0165-7836(00)00306-4.

Walsh, W.A., Kleiber, P., McCracken, M., 2002. Comparison of logbook reports of incidental blue shark catch rates by Hawaii-based longline vessels to fishery observer data by application of a generalized additive model. Fish. Res. 58, 79–94. https://doi.org/10.1016/S0165-7836(01)00361-7.

Table 1. Summary of model selection for binomial model. Δ AIC denotes a difference between AIC and a minimum value of AIC in all AICs.

Model	Explanatory variables	Number of parameters	Deviance	Δ AIC
M-1	Null	1	8894	535
M-2	Year	26	8522	212
M-3	Year, Month	37	8379	92
M-4	Year, Month, Lat15	39	8363	80
M-5	Year, Month, Lat15, Lon15	47	8285	18
M-6	Year, Month, Lat15, Lon15, HBF	50	8261	0

Table 2. Type-II analysis of deviance table components produced by binomial model. LR Chisq denotes Likelihood Ratio Chi-Square statistics, DF is degree of freedom, and Pr is significant probability for each factor.

Factor	LR Chisq	Df	Pr(>Chisq)
Year	422.75	25	< 0.001
Month	64.86	11	< 0.001
Latitude by 5^o	14.4	2	< 0.001
Longitude by 5^o	78.26	8	< 0.001
Hooks between float	23.68	3	< 0.001

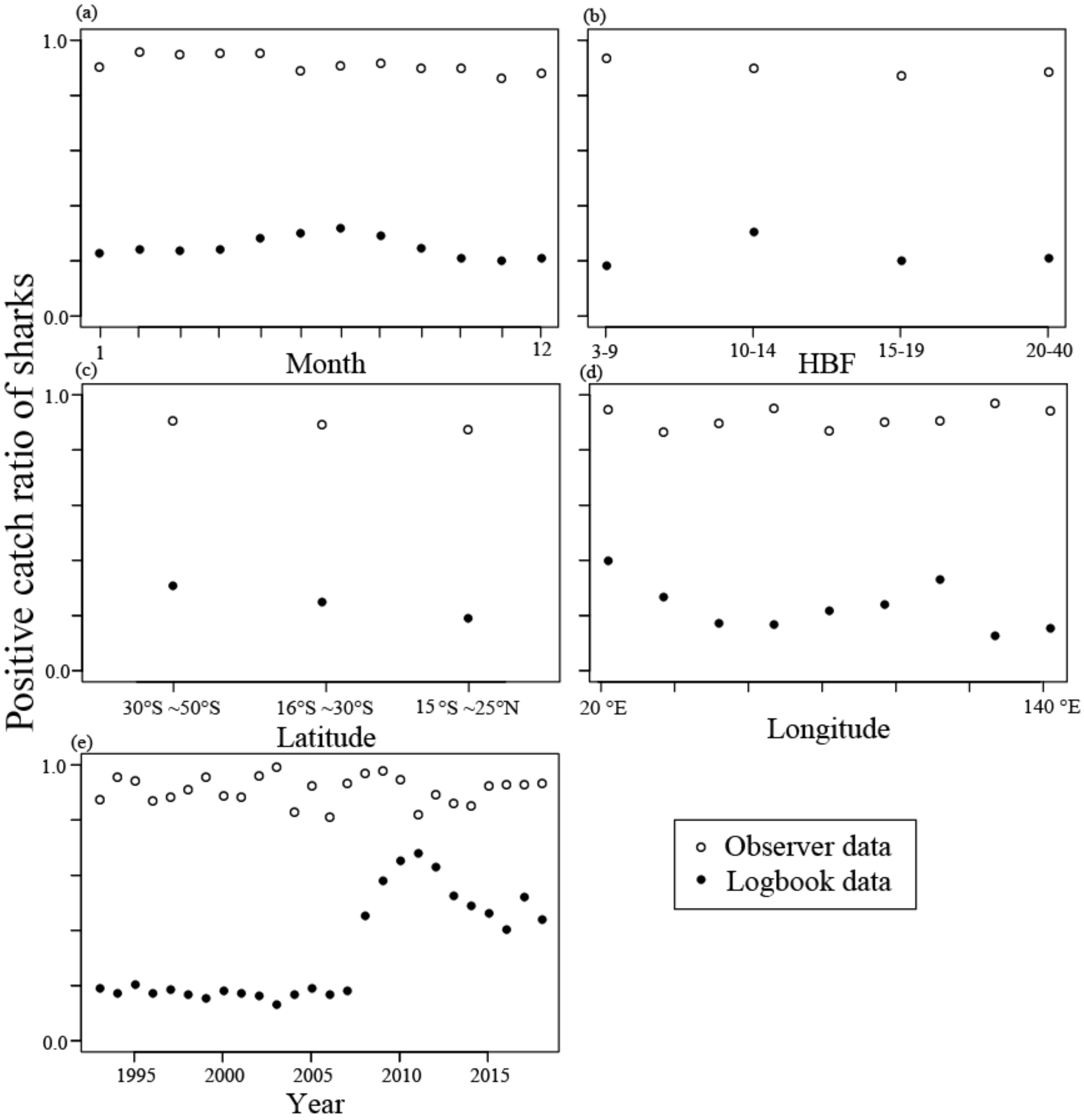


Fig. 1. Positive catch ratio of sharks (i.e. shortfin mako, blue shark, porbeagle shark, silky shark, oceanic white tip, hammerhead, and other sharks) for changes in values of each explanatory variable: (a) month; (b) hooks between floats (HBF); (c) latitude; (d) longitude and (e) year in binomial GLM.

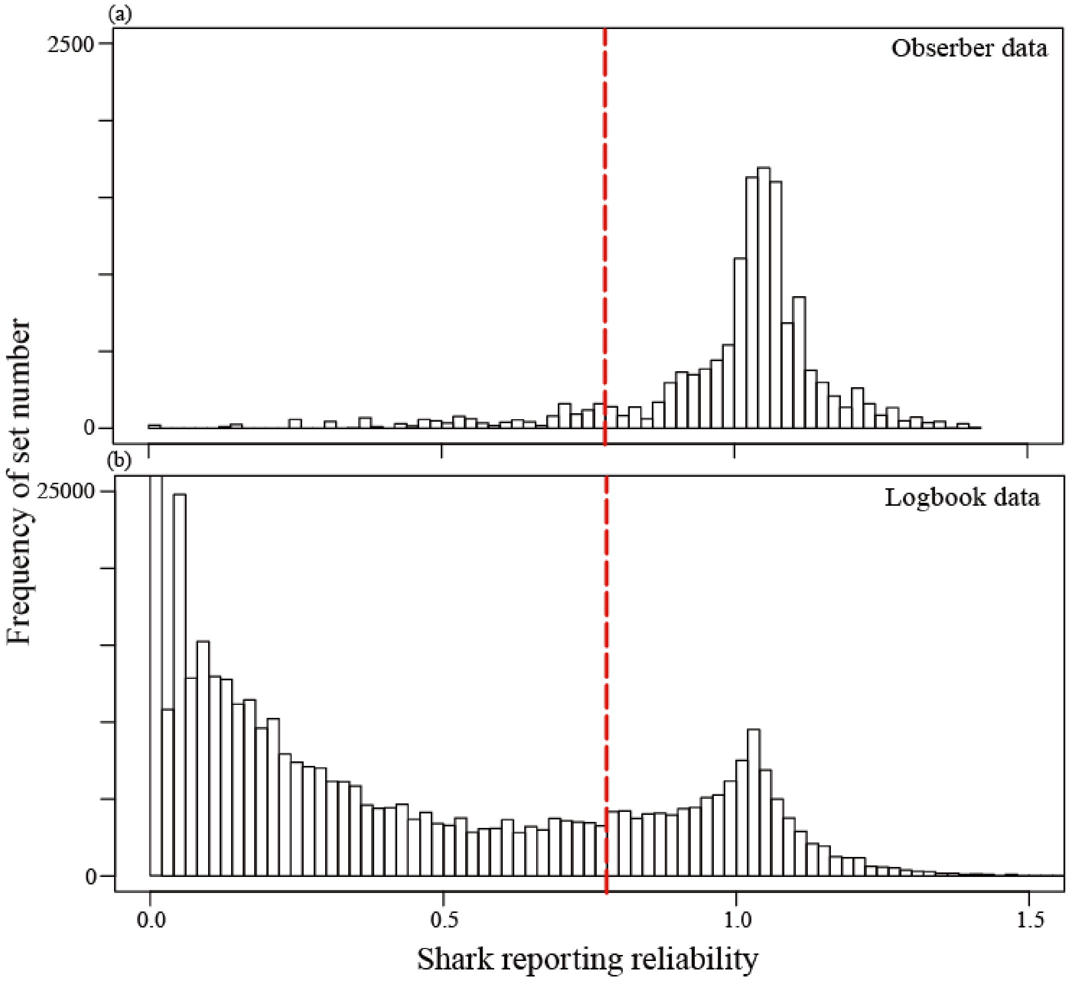


Fig. 2. Frequency of set number by shark reporting reliability computed from (a) observer data and (b) logbook data. Broken red vertical line denotes a threshold with lower bound of 80 % confidence intervals for data cut-off.

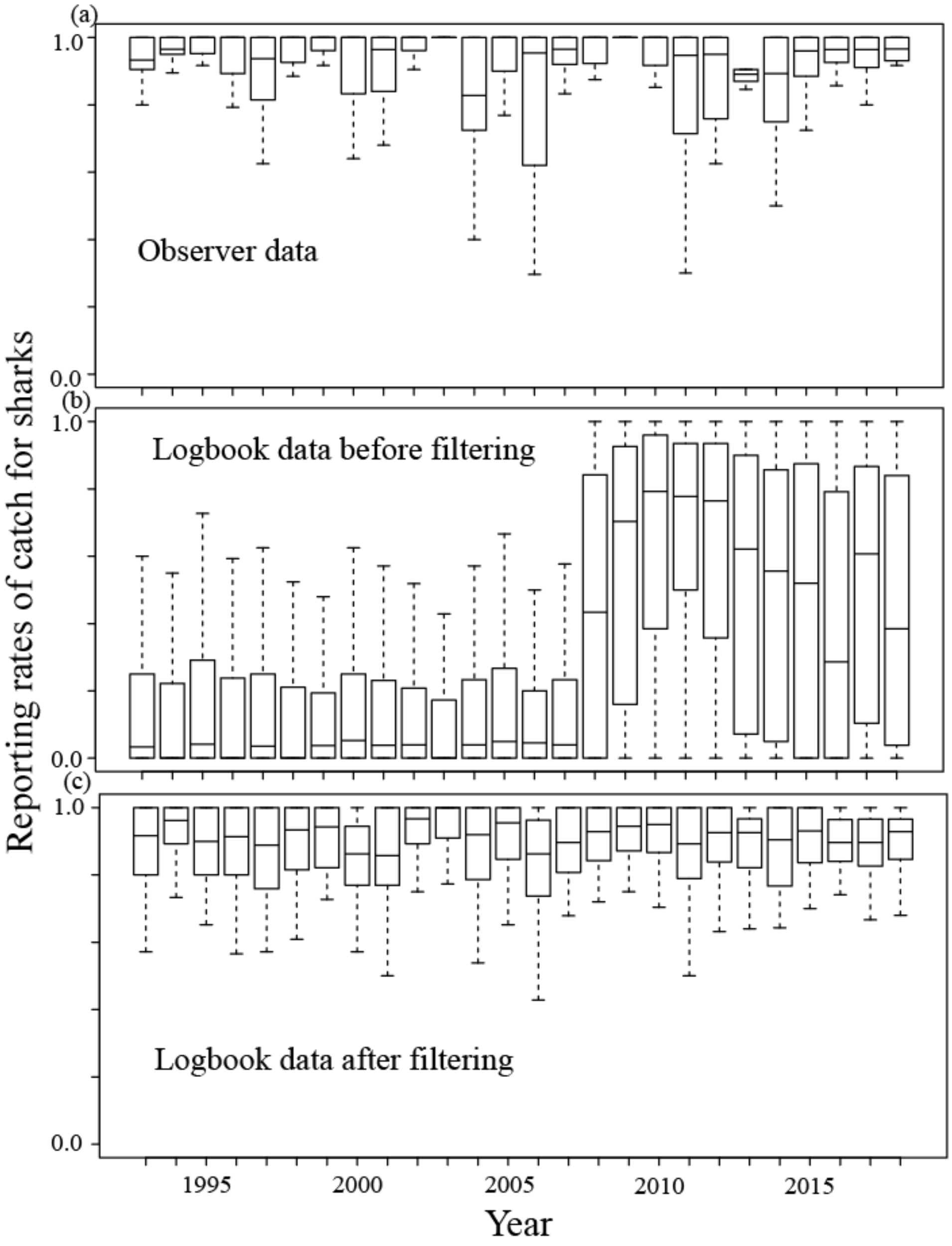


Fig. 3. Annual box plots of reporting rates of catch for sharks (a) observer data, (b) logbook data before filtering and (c) logbook data after filtering.

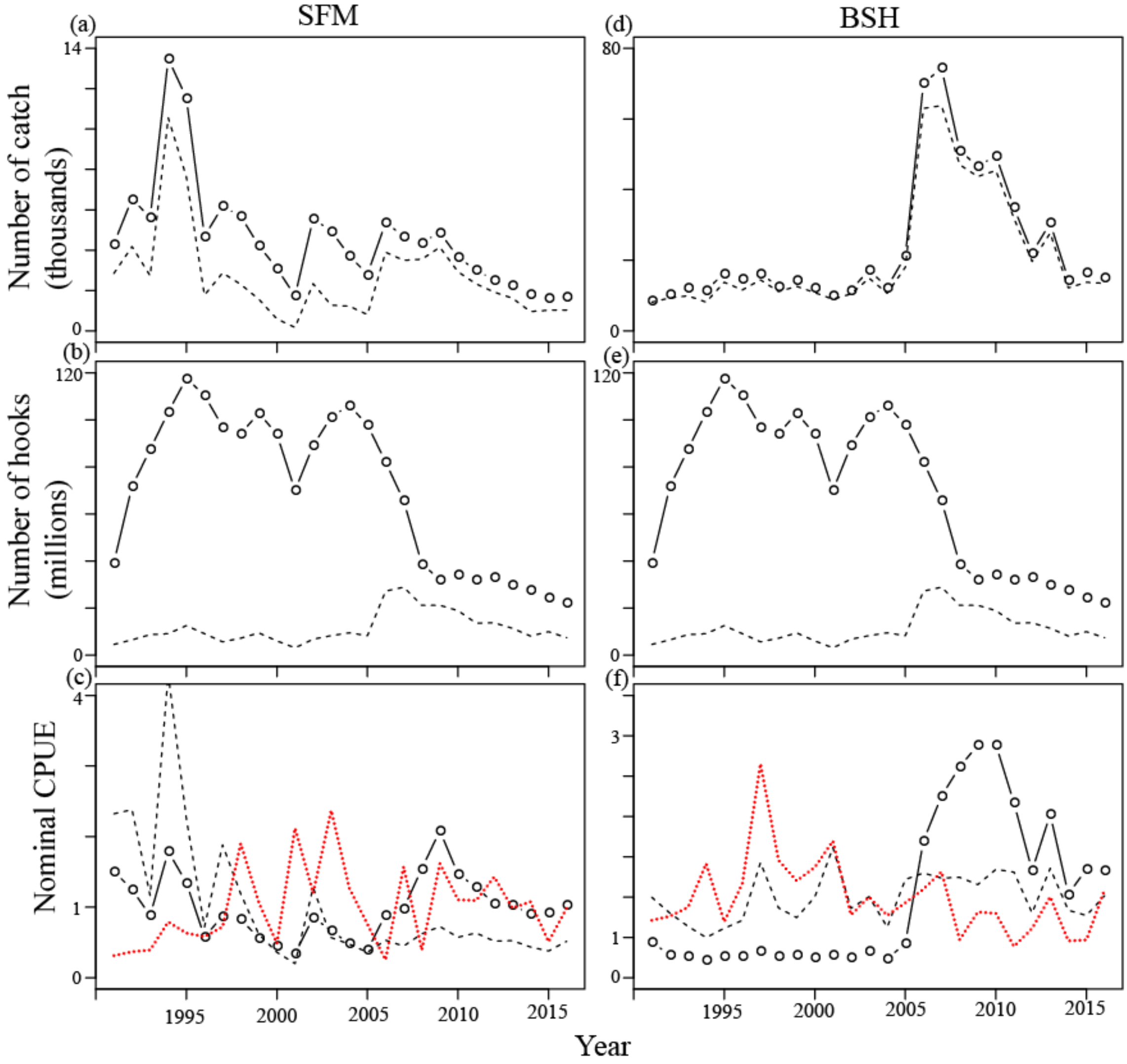


Fig. 4. Annual catch in numbers (thousands), number of hooks (millions), and nominal CPUE (per 1000 hooks) scaled by mean values for (a-c) shortfin mako (SFM) and (d-f) blue shark (BSH) for logbook data before filtering (solid line with open circles) and after filtering with 80 % data cut-off (broken line), and for observer data (dotted line with red color).

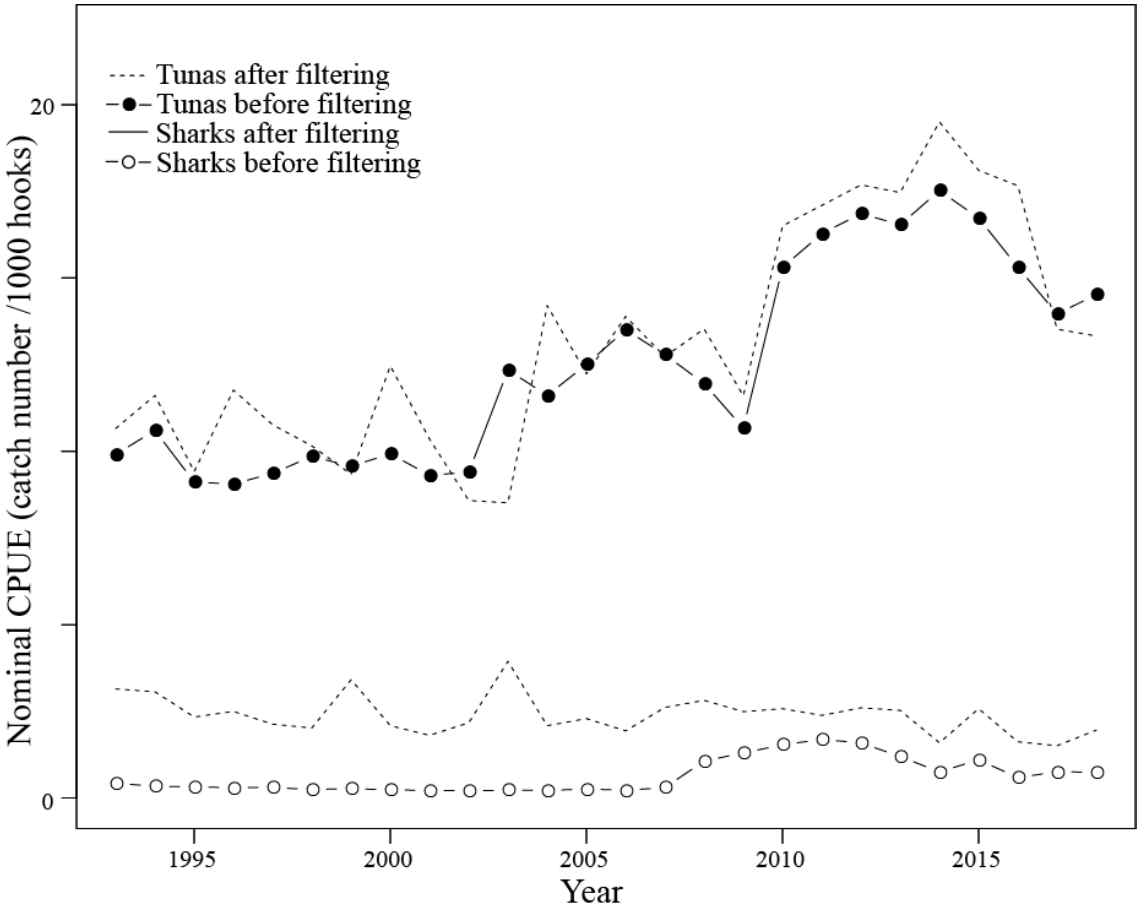
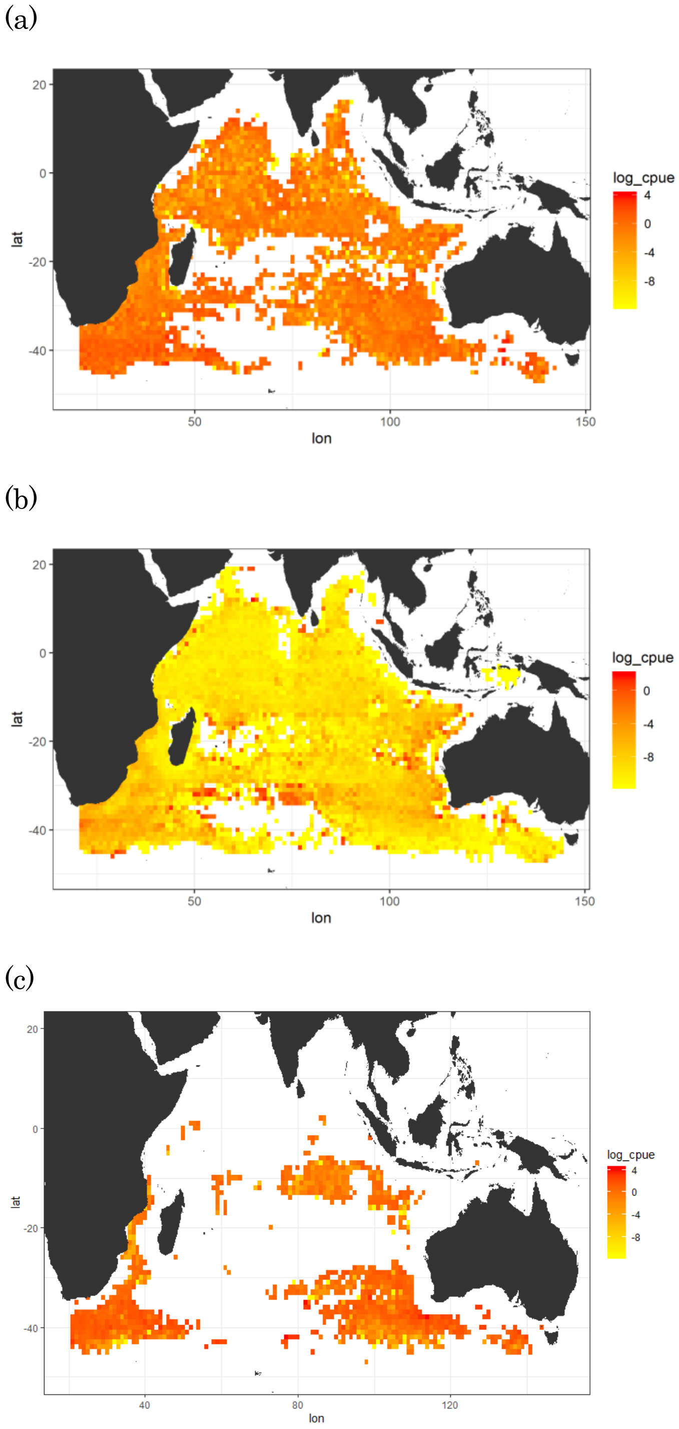


Fig. 5. Annual nominal CPUE for tunas and sharks before filtering (solid line with open circles) and after filtering (broken line).



(a)

(b)

(c)

Fig. A1. Overall spatial distributions of nominal log-CPUE for sharks in the Indian Ocean: (a) logbook data after filtering; (b)logbook data before filtering; (c)observer data. Data from 1993 to 2018 are combined.

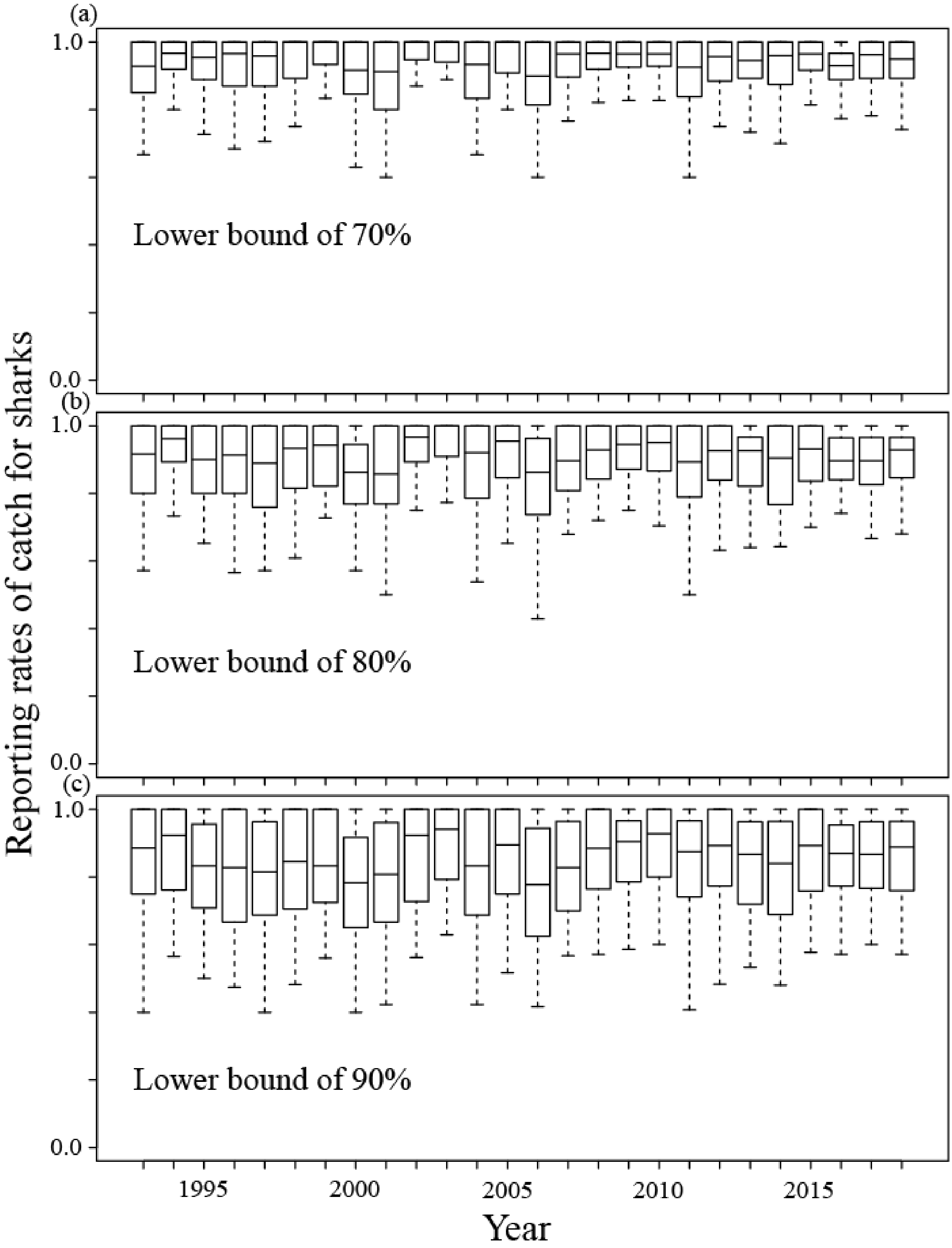


Fig. A2. Annual reporting rates of catch for sharks after filtering and cutting off the data with lower bound of (a) 70 %, (b) 80 %, and (c) 90% confidence intervals, respectively.

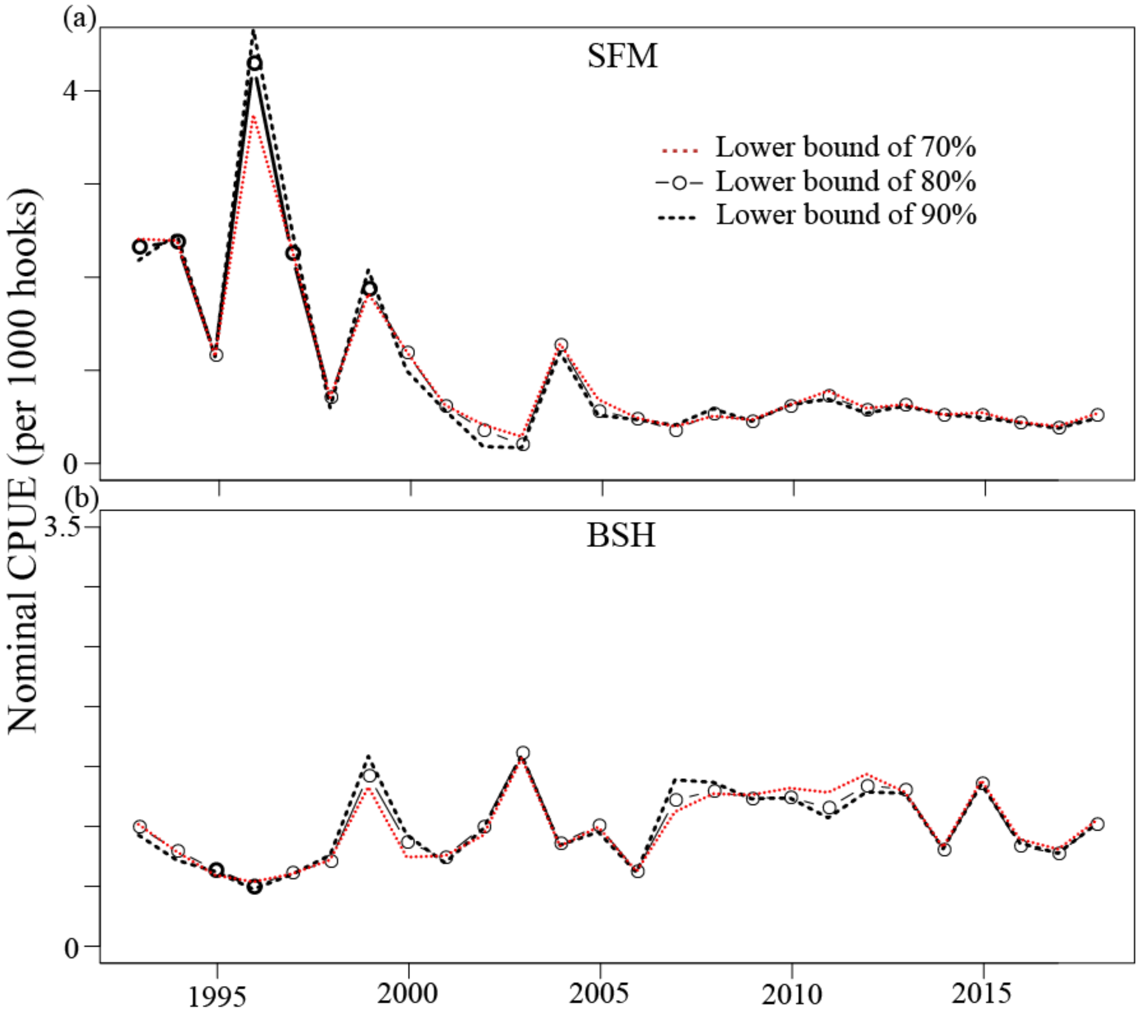


Fig. A3. Annual nominal CPUE (per 1000 hooks) scaled by mean values for (a) shortfin mako (SFM) and (b) blue shark (BSH) for logbook data after filtering with lower bound of 70 %, 80 %, and 90% confidence intervals, respectively.