ISC/22/ALBWG-02/04

CPUE standardization for North Pacific albacore caught by Japanese longline fishery from 1996 to 2021: the GLMM analysis using R-INLA

Jun Matsubayashi, Hirotaka Ijima, Naoto Matsubara Yoshinori Aoki and Yuichi Tsuda

Highly Migratory Resources Division, Fisheries Resources Institute,
Japan Fisheries Research and Education Agency (FRA)

2-12-4, Fukuura, Kanazawa-ku, Yokohama,
Kanagawa 236-8648, JAPAN

Email: Matsubayashi_jun86@fra.go.jp

This working paper was submitted to the ISC Albacore Working Group Intercessional Workshop, 6-12 December 2022, held at the Fisheries Resources Institute, Japan Fisheries Research and Education Agency, Kanagawa, Japan.

Summary

This study performed CPUE standardization of adult North Pacific albacore based on Japanese longline fishery operational data using geostatistical model and compared the results with that of previous study using WAIC obtained from Bayesian estimation. The main difference between these models is that the previous study incorporated spatial and temporal effects into the model as random effects, whereas this study incorporated these effects by spatiotemporal models with the Stochastic Partial Differential Equations (SPDE) approach. These models were intended to model albacore catch using year effect, location effect, hooks per basket, fleet type and vessel name. The results of model selection revealed that the application of SPDE significantly improves the performance of model (WAIC reduced by 63.6% in SPDE model) to standardize CPUE of albacore. In addition, we compared several models with different error distribution and with and without some explanatory variables (hooks per basket and fleet type) to search for the best model. The result of model selection showed that a spatiotemporal model with a zero-inflated negative binomial error distribution and incorporating all explanatory variables is the best model for CPUE standardization of adult North Pacific albacore.

Introduction

In North Pacific albacore (*Thunnus alalunga*) stock assessments, stock abundance indices (i.e., standardized CPUE) based on catch data from longline fisheries have been used as input data for stock synthesize models (ISC 2020). The standardized CPUE has been estimated by GLMM, which incorporates time and space effects as random effects (Ochi et al. 2017, Fujioka et al. 2019). However, this method ignores the fact that data closer in time and distance will have a greater correlation with the number of catches. For this reason, CPUE standardization using geostatistical models that account for spatial autocorrelation using the Stochastic Partial Differential Equations (SPDE) approach have recently been applied in the field of fisheries resource management (Ijima and Koike 2021). Thus, this study aims to compare the model for CPUE standardization used in the previous stock assessment of albacore (Fujioka et al. 2019) with standardized CPUE estimated by geostatistical model.

Commonly used spatio-temporal models in fisheries science that implement the SPDE approach are Integrated Nested Laplace Approximation (INLA; Rue et al. 2019) and Vector Autoregressive Spatio-Temporal Model (VAST; Thorson 2019). These models differ in terms of what criteria can be calculated as indicators of the model's performance; VAST can only calculate Akaike Information Criterion (AIC), whereas INLA can calculate various indicators such as Widely Applicable Information Criterion (WAIC;

Watanabe and Opper 2010) and Leave-One-Out Cross Validation (LOOCV). AIC is a commonly used indicator for model selection when using GLMM, but it does not work for complex models with hierarchical structure like spatio-temporal model (Watanabe and Opper 2010). On the other hand, WAIC was developed to compensate for the weakness of AIC and can be applied to even the most complex models. Since one of the main objectives of this study is to compare the model for CPUE standardization used in past resource assessments (e.g. Ochi et al. 2017, Fujioka et al 2019) with that of geostatistical models, we conducted our analysis using INLA, which is capable of calculating WAIC. Then, we provide annual trends of the standardized CPUE using the best fit model for potential use as input data for stock assessment model.

Data and Methods

Longline logbook data

The dataset for longline operations includes the number of albacores caught in each operation, date, quarter, fleet location type (Distant, Offshore, Coastal; hereafter *fleet*), hooks per basket (*hpb*), totals hooks and vessel ID from 1976 to 2021. The latitude and longitude of all data were recorded in units of one degree, and data taken in the same year, month, vessel, hpb, and latitude and longitude were summed up to reduce the amount of data. In order to focus on albacore fishery, data where the fleet location type was Distant and hooks per basket is smaller than 10 were preliminary excluded. Previous studies found that Area 2 (see Fig. 1) had larger fish (adults) regardless of season based on the catch at length data (Ijima et al., 2017) out of 5 main fishing Areas of albacore in the north Pacific (Ochi et al. 2016). In addition, the method of collecting logbook data for Japanese longline has changed since 1994, and stable data with the new collection method are considered to have been available since approximately 1996 (Ijima et al. 2017; Ochi et al., 2017; Fujioka et al., 2019; ISC 2019). For this reason , data from Area 2 and Quarter 1 since 1996 were used fort CPUE standardization to extract data most represent abundance of adult albacore.

Generation of INLA mesh

First, we converted the location of each data presented in latitude and longitude units into degrees from meter unit so that the distance of each data would be correctly reflected in the analysis. In order to perform modeling with INLA, it is first necessary to generate a mesh to create an artificial neighborhood set on the study area and calculate the spatial autocorrelation between data points. In creating the mesh, we must set the max.edge and cutoff values; the max.edge determines the largest allowed triangle length of the mesh

and the cuttoff value defines the minimum allowable distance between points. Higher resolution is obtained with lower max.edge value, whereas that increase the computational time. Thus, the value of max.edge is determined by a trade-off with the computation time. Since points within the cutoff value are replaced by a single vertex prior to the mesh refinement step, smaller cutoff value yields higher-resolution mesh, but it also needs to be set in a trade-off with the computation time. In this study, the mesh was created with a max.edge of 500 and cutoff value of 170 (Fig. 2).

Reconstruction of iid model

The model used by Fujioka et al. (2019) in the previous stock assessment incorporated the effect of location as independent random effect (iid). For this reason, we refer to this model as *iid* model. The model structure of *iid* model was as follows:

$$CPUE_{alb} \sim intercept + year + fleet + hpb + f(vessel\ ID, model = iid) + f(latlon, model = iid) + offset(hooks/1000)$$

where $CPUE_{alb}$ is estimated standardized CPUE for albacore, $latlon$ is the location of operation rounded to the nearest 5 degrees of latitude and longitude and $model = iid$ represents variable incorporated as random effect. All explanatory variables were incorporated into the model as categorical variables. The response variable was assumed to follow negative binomial distribution, and the number of hooks divided by 1000 was used as an offset term to unify the amount of effort. Although Fujioka et al. (2019) constructed the model using RSTAN package (Stan Development Team 2022) in R, this study implemented the *iid* model using R-INLA to calculate WAIC and LOOCV.

Generation of SPDE models

Unlike the *iid* model, geostatistical model treats the effect of location through the SPDE approach, and thus we refer to this model as *SPDE* model. The structure of *SPDE* model with all explanatory variables (full model) is as follows:

$$CPUE_{alb} \sim intercept + year + f(fleet, model = iid) + f(hpb, model = iid) + f(vessel\ ID, model = iid) + f(w, model = AR1) + offset(hooks/1000)$$

where w is the spatial random effect calculated based on the SPDE approach and $AR1$ represents autoregressive model. Thus, the model estimates multiple spatial random fields

that are autoregressive by year. Unlike the *iid* model, $fleet$ and hpb were incorporated as random effect for technical reason, but the difference influence on neither the WAIC nor model selection. The other parts of the model are the same with *iid* model.

In addition to the full model described above, models combining different explanatory variables and error distribution of response variable were compared to search for the best model. Given the long computation time required to implement the spatio-temporal model, the candidate explanatory variables to be excluded were *fleet* and *hpb*, and a negative binomial distribution was tested in addition to the zero-inflated negative binomial distribution as the error distribution of the response variable.

Results and Discussion

Model selection

The performances of *iid* and *SPDE* models were compared based on WAIC and LOOCV of full models. The WAIC and LOOCV for *iid* model were 1771956.0 and 1436367.0, respectively, whereas that of *SPDE* model were 642479.4 and 642346.4, respectively. Thus, the *SPDE* model showed better performance than *iid* model, and the improvement of WAIC was no less than 63.6%. Results of comparison of the SPDE models assuming different error distributions for the response variable and combinations of explanatory variables, zero-inflated negative binomial distribution incorporating all explanatory variables had the lowest WAIC, although differences in WAIC among SPDE models were very low (Table 1). Thus, we concluded that *SPDE* model with zero-inflated negative binomial distribution including all explanatory variables was the best model to estimate standardized CPUE for adult north Pacific albacore.

Validation of the best model

Plot of Matérn correlation function, which defines the correlation between locations, versus distance suggested that we have strong spatial correlation up to about 300km, and the distance over which the correlation between points decreases to 10% was approximately 880 km (Fig. 3). Therefore, the max.edge value we set for the triangularization was within the range of distances where the correlations between sites were sufficiently high.

Plot of randomized quantile residuals suggested that the residuals generally followed a normal distribution (Fig. 4), but for the points where the residuals were less than -2, there may still be some systematic error as clearly did not follow a normal distribution. Negative residuals indicate that actual catches were lower than expected. To investigate the causes of these systematic errors, we checked the relationships between randomized

quantile residuals and year, location, hpb, and fleet, but no clear differences along with these variables were found. The only residual plot by fishing vessel ID showed residuals that clearly deviated from the normal distribution for certain vessels. Even for the same fishing vessel, the residuals sometimes follow a normal distribution and sometimes do not, and it is difficult to identify the cause of these systematic errors. However, such vessels may not have achieved the expected catches due to lack of technology or targeting species other than albacore. In future stock assessments, more accurate standardized CPUE could be achieved by introducing a process to extract and exclude such problematic data. No other problems were found with the latent random field and the posterior distribution of parameters of the best model (Fig. 5, 6).

Estimation of standardized CPUE

The trends in annual changes in standardized CPUE estimated based on the best model were generally consistent with the trends in nominal CPUE, whereas standardized CPUE tended to be lower than the nominal CPUE after 2000 (Fig. 7). On the other hand, the standardized CPUE estimated by using the *iid* model (Fig. 8; Matsubayashi et al. 2022) did not show such trend, and the standardized CPUE was always close to the nominal CPUE. Given the significant improvement in WAIC in the *SPDE* model, we recommend using the standardized CPUE estimated by INLA presented in this study as the input data for the stock assessment model.

Reference

Fujioka, K., Ochi, D., Ijima, H., Kiyofuji, H. (2019) Update standardized CPUE for North Pacific albacore caught by the Japanese longline data from 1976 to 2018. ISC/19/ALBWG-02/01.

Ijima, H., Ochi, D. and Kiyofuji, H. (2017) Estimation for Japanese catch at length data of North Pacific albacore tuna (*Thunnus alalunga*). ISC/17/ALBWG/04.

Ijima, H. and Koike, H. 2021. CPUE standardization for Striped Marlin (*Kajikia audax*) using spatio-temporal model using INLA. ISC/21/BILLWG-03/01.

ISC (2020) Stock assessment of albacore tuna in the north Pacific Ocean in 2020. Report of the albacore working group, 15–20 July 2020, online.

Ochi, D., Ijima, H., Kinoshita, J. and Kiyofuji, H. (2016) New fisheries definition from Japanese longline North Pacific albacore size data. ISC/16/ALBWG-02/03.

Ochi, D., Ijima, H. and Kiyofuji, H. (2017) Abundance indices of albacore caught by Japanese longline vessels in the North Pacific during 1976-2015. ISC/17/ALBWG/01.

Matsubayashi, J., Ijima, H., Matsubara, N., Aoki, Y., and Tsuda, Y. 2022. Updated CPUE standardization for adult North Pacific albacore caught by Japanese longline fishery from 1996 to 2021: the GLMM analysis using STAN. ISC/22/ALBWG-XX/XX.

Rue H., Martino S., Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B, 71, 319–392.

Stan Development Team (2022) RStan: the R interface to Stan. R package version 2.21.7, https://mc-stan.org/.

Thorson J.T. (2019) Guidance for decisions using the vector autoregressive spatio-temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fisheries Research, 210, 143–161.

Watanabe, S. and Opper, M. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of machine learning research, 11, 3571–3594.

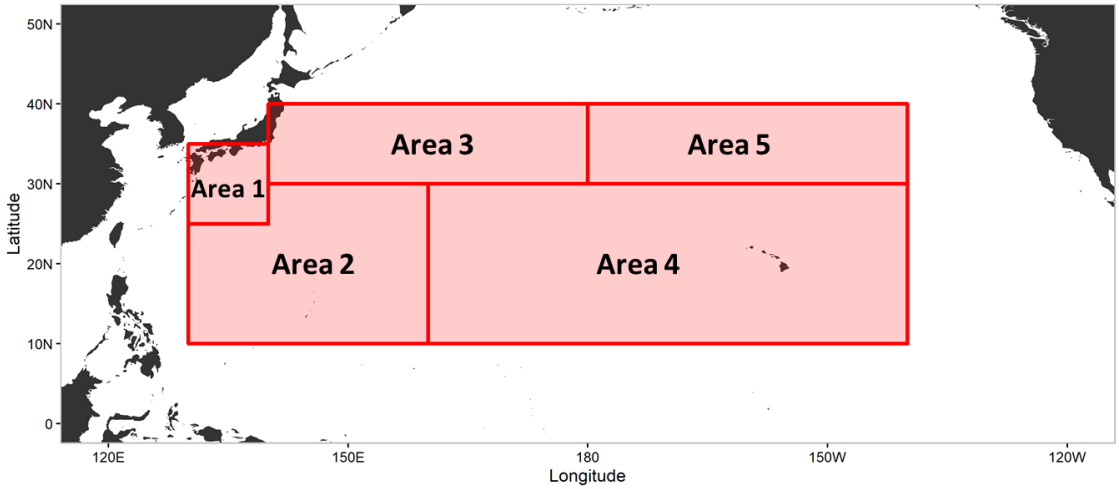

Figure 1. Area definition of Japanese longline fishery for albacore.

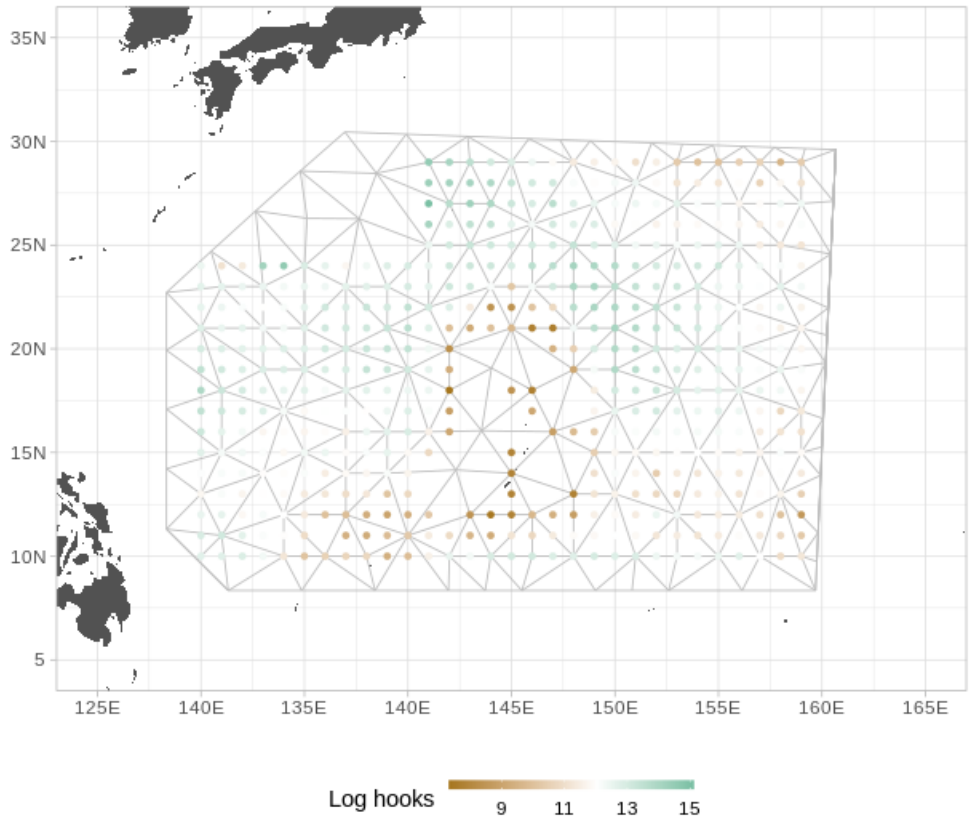

Figure 2. Triangularization of Japanese longline data and points are data locations. The color of each point indicates the number of hooks.

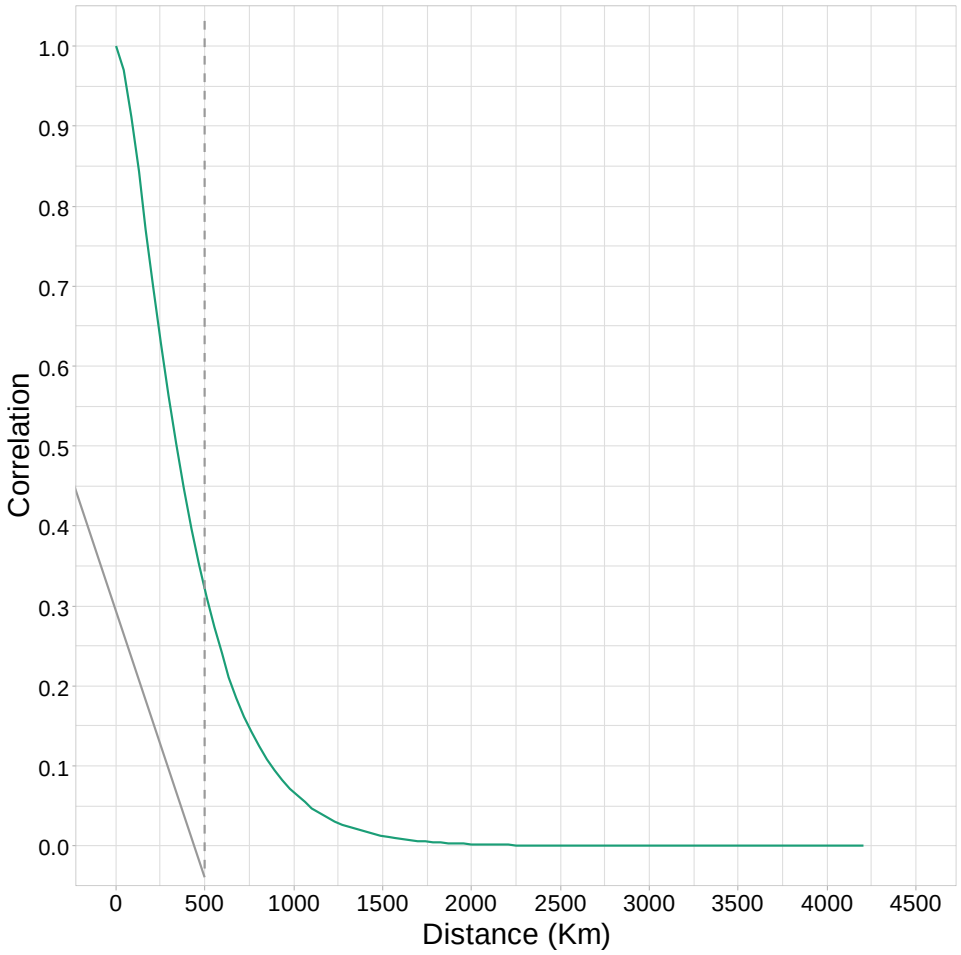

Figure 3. Matern correlation function versus distance. The dashed line indicates the max.edge value we set for triangularization.

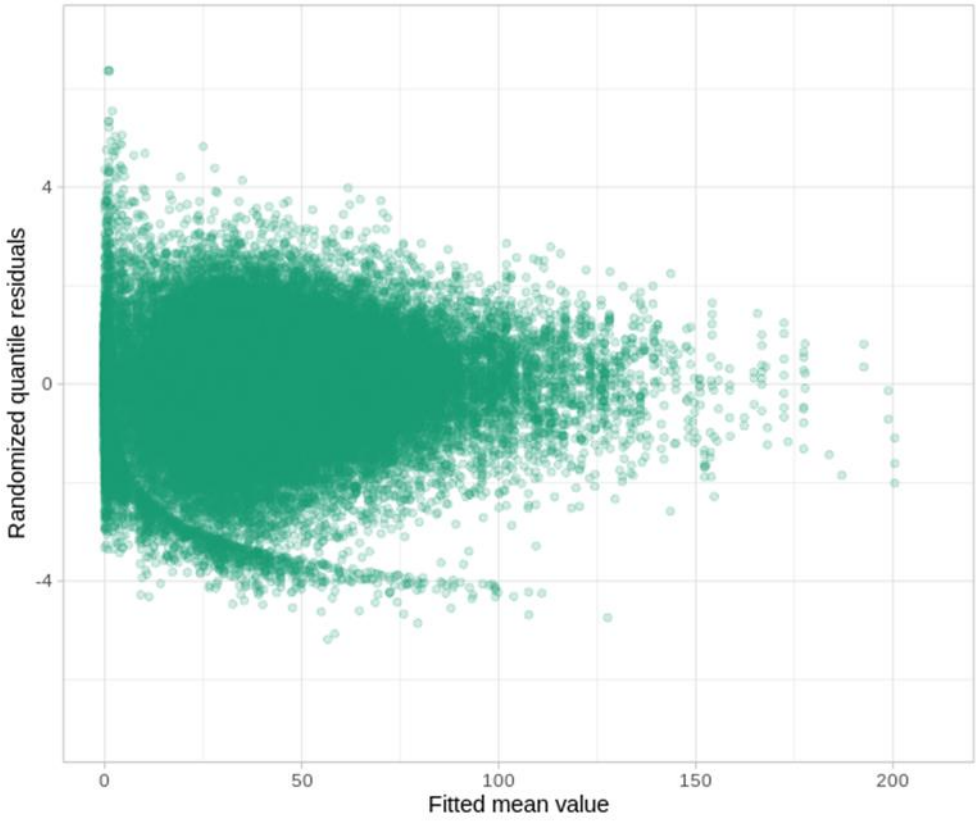

Figure 4. Plot of randomized quantile residuals versus fitted values.

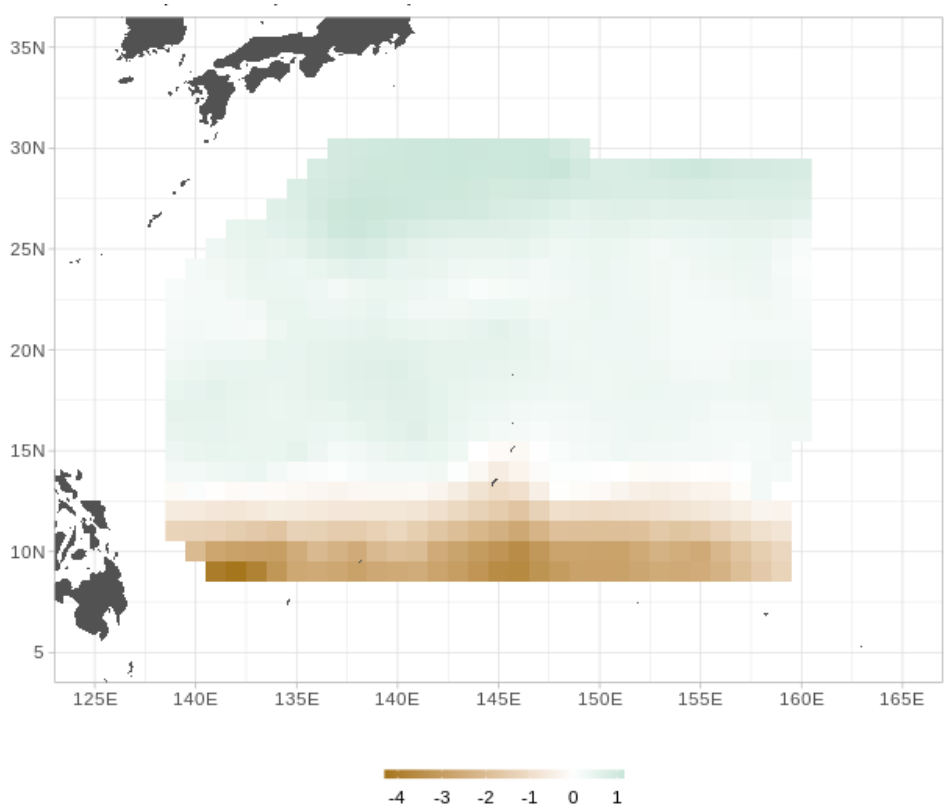

Figure 5. Interpolated spatial random field (annual mean values are shown).

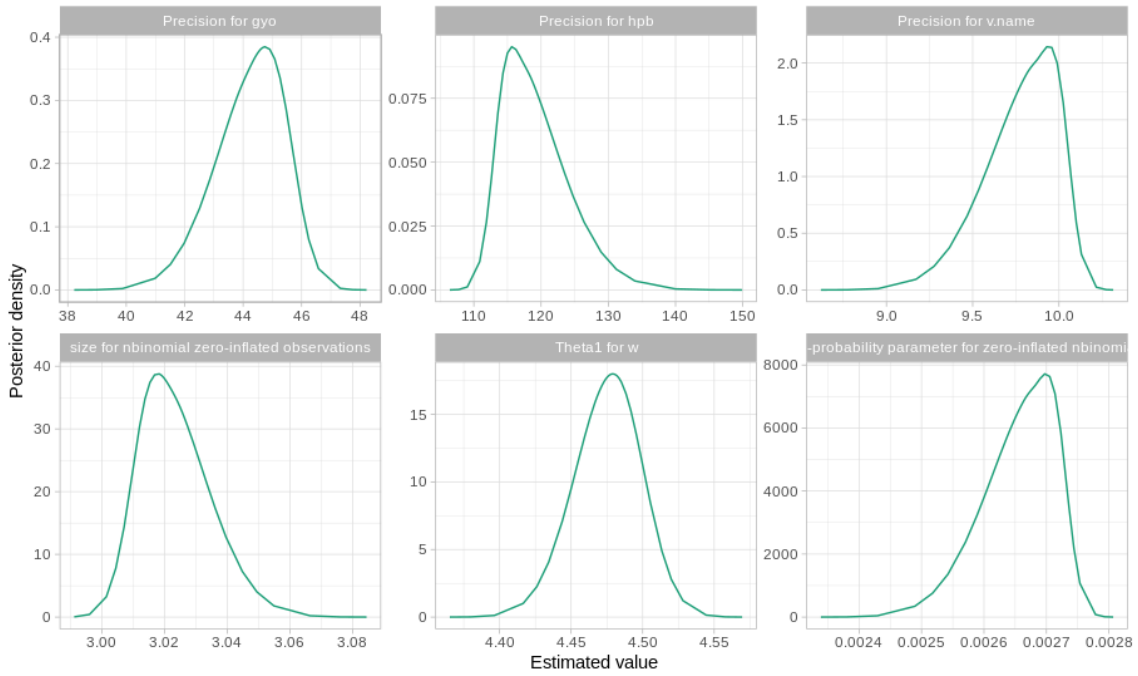

Figure 6. Posterior distributions of each parameter.

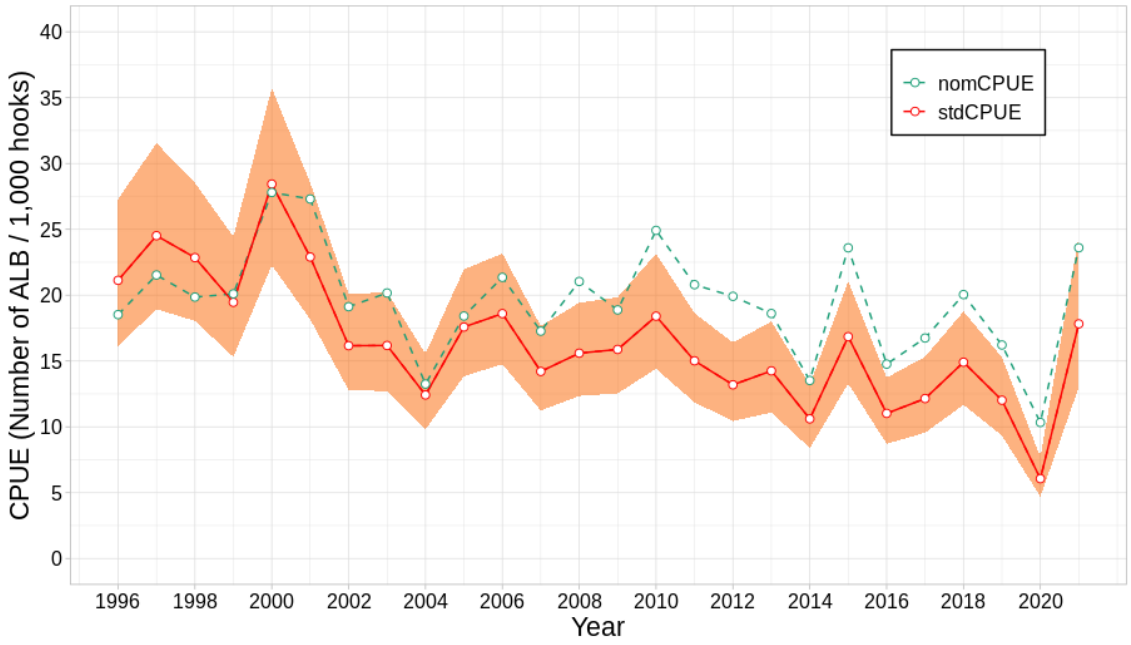

Figure 7. Plot of annual trends in nominal and standardized CPUE estimated by using SPDE model in this study. The red ranges indicate the 5% and 95% quantile intervals of the estimated standardized CPUE.

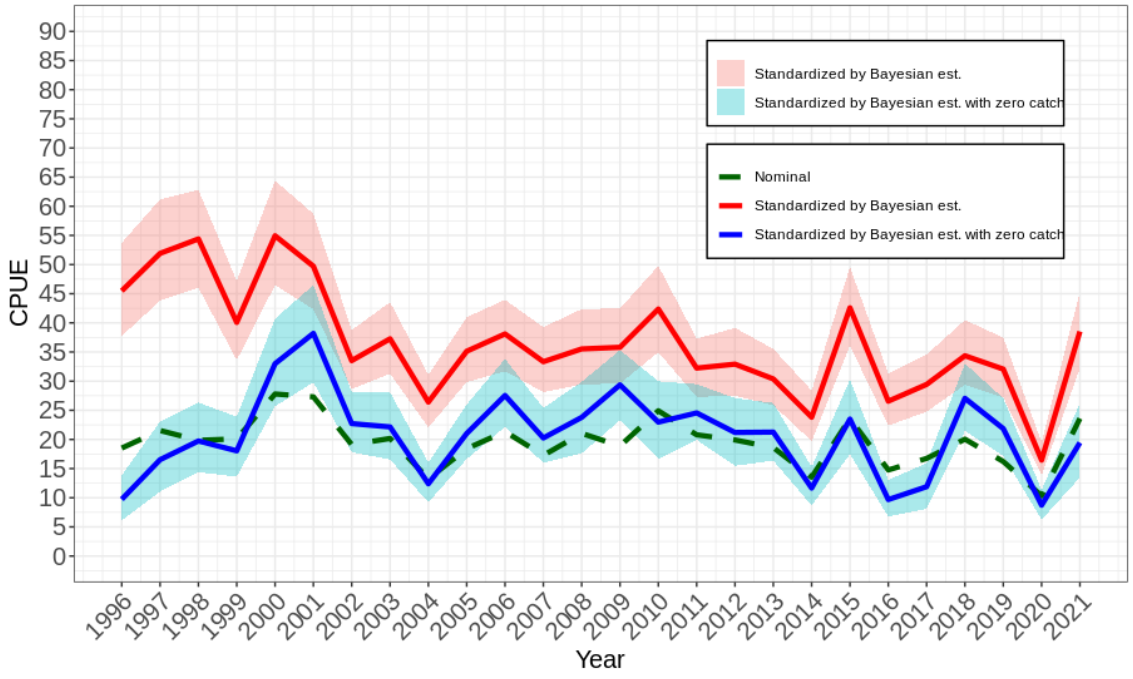

Figure 7. Plot of annual trends in nominal and standardized CPUE estimated by iid model in Matsubayashi et al. (2022). The red ranges indicate the 5% and 95% quantile intervals of the estimated standardized CPUE.

Table 1. Results of model selection. Lowest values of Widely Applicable Information Criterion (WAIC) and Leave-One-Out Cross Validation (LOOCV) were highlighted in red.

Model type	Family	Model structure (INLA function)	WAIC	LOOCV
iid model	Zero-inflated negative binomial	$CPUE_{alb}$ ~ *intercept* + *year* + *fleet* + *hpb* + *f*(*vessel ID*, model=*iid*)+ *f*(*latlon*, model=*iid*) + offset(*hooks*/1000)	1771956	1436367
SPDE model	Zero-inflated negative binomial	$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*fleet*, model=*iid*) + *f*(*hpb*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	642479.4	642346.4
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*hpb*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	644031.8	643755.4
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*fleet*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	644328.7	644006.8
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	644229.1	644004.1
	Negative binomial	$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*fleet*, model=*iid*) + *f*(*hpb*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	643790.8	643588.6
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*hpb*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	643801.8	643600.5
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*fleet*, model=*iid*) + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	643934.1	643706.9
		$CPUE_{alb}$ ~ *intercept* + *year* + *f*(*vessel ID*, model=*iid*)+ *f*(*w*, model=*AR1*) + offset(*hooks*/1000)	643954.2	643728.7