Size-dependent distribution of Pacific Striped Marlin (*Kajikia audax*) : The analysis of Japanese longline fishery logbook data using the finite mixture model. [1]

*Hirotaka Ijima and **Minoru Kanaiwa
E-mail:ijima@affrc.go.jp

*National Research Institute of Far Seas Fisheries, Fisheries Research and Education Agency
5-7-1, Orido, Shimizu, Shizuoka, 424-8633, Japan
**Mie University
1577, Kurima-Machiyacho, Tsu, Mie, 514-8507, Japan

[1]This working paper was submitted to the ISC Billfish Working Group Intercessional Workshop, 14-21 January 2019, held in the Pacific Islands Fisheries Science Center of the National Marine Fisheries Service, Hawaii USA.

Abstract

The size selectivity by fishing gear and CPUE are the essential information configuring the integrated stock assessment models such as the stock synthesis 3. However, the selectivity and CPUE need to define by time-spatially because the fishing operation corresponds to fish life-history such as growth and migration or distribution. Here, to clarify the spatial distribution pattern of the Pacific striped marlin (*Kajikia audax*), we explored the finite mixture model analysis using the R software package "flexmix". In this analysis, we used the Japanese longline operational data that reports operating area, catch number, effort (number of hooks) and catch weight. Regarding model assumption, we set 2 to 12 clusters with two-dimensional GLMs that responses are mean body weight and CPUE. We used several covariates for these GLMs (e.g., year, quarter and gear effects). We also set the area-seasonal grouping factor. We used BIC for the model selection and, BIC selected complex nine cluster model for the entire Pacific. Considering with spatial cluster and trends of mean-body weight, we suggest 11 Japanese longline fleets for the model configuration of Stock synthesis 3.

Introduction

ISC Billfish working group used the Stock Synthesis 3 (SS3) for the stock assessment of striped marlin in the Western and the Central North Pacific Ocean (WCNPO) (ISC 2015). The SS3 needs to define the fleets that depend on fishing size selectivity. Usually, the size selectivity is estimated by size composition data, then SS3 calculate catch at length. Thus, determining the size selectivity is one of the most critical configurations for SS3. However, estimated selectivity was sensitive in the previous striped marlin stock assessment because the length frequency data changed time-spatially that depends on fish life history (e.g., growth and migration or distribution)(ISC 2015).

SS3 can consider the historical change of selectivity in the parameter estimation, but to sets the spatial difference is difficult. Therefore BILLWG needs to define the area dependent selectivity outside of the SS3 (Waterhouse, Sampson, Maunder, and Semmens 2014). Some research focuses on area-based size selectivity, and they applied the cluster analysis or the generalized additive model (GAM) using length frequency data (Ochi, Ijima, Kinoshita, and Kiyofuji 2016), (Langseth 2016). In the previous stock assessment, Japanese longline fleet definition considered size distribution, however any statistical analysis was not addressed. On the other hand, Catch pre unit effort (CPUE) also needs to divide by area because CPUE represents trends of biomass and also changed by time-spatially. Thus, it needs size information and CPUE simultaneously to determine the accurate fleet definition of SS3. The finite mixture model (Leisch 2004) is useful to explore such difficulties for WCNPO stock assessment because 1) the finite mixture model can divide mixed distribution, 2) the finite mixture model can assume two kinds of response variables. Thus, using the finite mixture model, we can address statistical analysis using the size and CPUE information simultaneously (Ijima and Kanaiwa 2018), (Kinoshita, Aoki, Ijima, and Kiyofuji 2018).

Here, we defined the Japanese longline fishery fleet for SS3 using finite mixture model with two response variables as mean body weight and CPUE. In the finite mixture model analysis, we analyzed an entire Pacific pattern of striped marlin then focus on WCNPO.

Material and methods

Datasets

We used Japanese longline logbook data rather than length frequency data for finite mixture model analysis. The advantage of logbook data are 1) high area resolution, 2) longer time series

(1994-2017) and 3) low sampling bias. For example, Japanese logbook data can combine length frequency data but, sampling was biased by area and vessel (commercial and training vessel) and available time series are short (1999-2017).

Logbook data describes trip ID, vessel name, operational date, operational location, number of striped marlin catch, and catch amount of striped marlin. The pattern of striped marlin CPUE and mean catch weight change by season and mean catch weight is different by area (Figure.1). We aggregated logbook data by year, quarter, 5 ° × 5 ° area, gear (hooks between floats, shallow: < 8, deep: >= 8). Then we removed zero catch rows, because weight zero fish is unrealistic.

Finite Mixture Model

The log-likelihood for N independent observations of the Finite Mixture Model with K clusters and D-dimensional response variables $\boldsymbol{y} = (y_1, ..., y_D)'$ are

$$\log L = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} f_{d,k} \left(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta}_{k,d} \right) \right), \tag{1}$$

and

$$\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1. \tag{2}$$

Where $\boldsymbol{x}$ is an independent variables vector, π_k is the prior probability of the cluster k , and $\boldsymbol{\theta}_{k,d}$ is the clusters with the dimension specific parameter vector of the density function f_d. Flexmix use Expectation-maximization (EM) algorithm for Maximum-likelihood estimation.

In this analysis, we used two Generalized Linear Models (GLMs) as the density function f_d $(d = 1(weight), 2(cpue))$ that responses are mean body weight (W) by one operation and CPUE (C). Firstly, we assumed log normal density GLM (f_1) for mean body weight as follows:

$$\begin{aligned} \boldsymbol{\theta}_{k,1} &= \boldsymbol{b}_{k,1}, \boldsymbol{a}_1, \sigma^2_{k,1} \\ \log(W_k) &\sim Normal(\mu_{k,1}, \sigma^2_{k,1}) \\ E(\log(W_k)) &= \mu_{k,1}, var(\log(W_k)) = \sigma^2_{k,1} \\ \log(W_k) &= \boldsymbol{X}_k \boldsymbol{b}_{k,1} + \boldsymbol{Z}_k \boldsymbol{a}_1. \end{aligned} \tag{3}$$

Where $\mu_{k,1}$ is the mean of normal distribution, $\sigma^2_{k,1}$ is the variance of normal distribution, W_k is the response vector of the individual mean body weight, $\boldsymbol{X}_k$ and $\boldsymbol{Z}_k$ are the variable matrix, $\boldsymbol{b}'_{k,1}$ and $\boldsymbol{a}'_1$ are the parameter vector in cluster k and parameter vector that is not changed by throughout the cluster. variance $\sigma^2_{k,1}$ is scalar in cluster k. We assumed the variable as year, quarter and gear (hooks between floats) and all variables are treated as the categorical.

Secondly, we constructed poisson GLM (f_2) for CPUE is

$$\begin{aligned} \boldsymbol{\theta}_{k,2} &= \boldsymbol{b}_{k,2}, \boldsymbol{a}_2, \\ C_k &\sim Poisson(\mu_k), \\ E(C_k) &= var(C_k) = (\mu_k), \\ \log(\mu_k) &= \boldsymbol{X}_k \boldsymbol{b}_{k,2} + \boldsymbol{Z}_k \boldsymbol{a}_2 + \log(1000 hooks_k). \end{aligned} \tag{4}$$

Where $\mu_{k,1}$ is the mean and variance of the poisson distribution, C_k is the response vector of CPUE, $\boldsymbol{X}_k$ and $\boldsymbol{Z}_k$ are the variable matrix, $\boldsymbol{b}_{k,2}$ and $\boldsymbol{a}_2$ are the parameter vector in the cluster k and parameter vector that is not changed by throughout the cluster. $\log(1000 hooks)$ are offset variable. We assumed variables as year, quarter and gear.

We made six models that were considered different assumption (Table.1). Area variable ($5\ ^\circ \times 5\ ^\circ$ grid data) or area-seasonal variable was set to grouping factor (Table.1) because, our object is to define area dependent fishery definition for the SS3. All parameters were estimated by R software package "flexmix" ver2.3-14.

Model selection and validation

To chose the appropriate number of area cluster, we set 2-12 clusters for the prior cluster on the six flexmix models. We use a Bayesian information criterion (BIC) for the model selection. In the model validation, we confirmed 1) posterior probabilities of the estimated cluster, 2) Pearson residuals of two GLM models, and 3) coefficients significance of GLM. Finally, we plotted estimated clusters spatially and compared with the spatial trend of CPUE and weight of individual fish.

Result and discussion

Model selection and validation

As a result of model selection, BIC selected m2rev1with nine clusters (Table.2 ,Figure.2). This model set ten clusters as a prior value, and "flexmix" estimated nine clusters. On the modeling process, we modified m2 model to m2rev1 model. At first, the smallest BIC was obtained by the m2 model. However, most of the parameters of the year effect estimated in the mean body weight GLM were not statistically significant. Therefore, we removed the covariate of the year effect from the mean body weight GLM and analyzed again. The two GLM of m2rev1 estimated a total of 261 parameters, and most of which (245) was statistically significant ($p < 0.05$). This result suggests that the catch size obtained from Japanese longline fishery changes in area and quarter, but indicate the effect by year is small. The posterior of all estimated data shows that 0 or 1 is relatively large in all clusters (Figure.3). That is, cluster estimation accuracy is relatively high. The fitness of two GLM of m2rev1 is also well because Pearson residuals of m2rev1 are almost evenly distributed well (Figure.4 - Figure.6).

However, there are rooms for improvement in this model. For example, this model needs to consider the random effect of the vessel name, because it is well known that the random effect of the vessels names has a considerable influence on the CPUE standardization (Ochi, Ijima, and Kiyofuji 2017). In this study, however, "flexmix" with the random effect GLM using R software package lme4 (Bates, Maechler, Bolker, Walker, Christensen, Singmann, Dai, and Grothendieck 2015) did not converge. There is also room for consideration about the handling of zero catches. In CPUE's GLM, zero catch can be considered, but there remains a problem on how to treat weight at zero catch in average weight GLM. In this study, we aggregated logbook data and removed zero catch rows but to delete data is not desirable.

The distribution pattern of Pacific striped marlin

The estimated nine clusters showed a different trend for each season (Figure.7). The pattern of the clusters are complicated, but cluster 5 and 6 spread across the north and south centered on the equator throughout the year (Figure.7). Especially cluster 6 has very low CPUE and divides the distribution of Pacific striped marlin in the north and south Pacific. Besides, the size of cluster 6 changed according to the season (Figure.7). Focusing on WCNPO, clusters other than cluster 6 show a mosaic pattern, and 0 to 3 age fish are mixed and distributed (Figure.8). However, we confirmed the cluster 5 of the first quarter where large fish appear, clusters 1 and 7 of the first quarter and cluster 1 of the fourth quarter where age 0 fish appears (Figure.8).

Considering these results, we propose 11 Japanese longline fleets for SS3 model configuration (Figure.9). In the first quarter, we divided into four areas. In Area 1, over three years old fish

are caught, Area 2 where one and two years old fish are caught. Age 0 fish has been caught a lot in the Area3. Area 4 is a low-density area. In the second and third quarter, we divided into two areas as low-density and well-mixed area (Area1 and Area2). In the fourth quarter, we classified three areas (Area1-Area3). One to two-year-old fish are caught in Area1, A lot of age 0 fish appears in Area 2, and density of striped marlin is low in the Area3.

Concerning CPUE standardization, we propose to standardize with two longline fleets that are Area1 of the first quarter and Area 1 of the third quarter. Since the clusters show in a mosaic pattern, Area 1 (<180 °E of the third quarter) and the first quarter of Area 1 where the large fish are caught are relatively coherent. According to the standardization of CPUE, it is also possible to increase the fleet definition in the third quarter to three.

Summary and suggestions

- Estimated clusters are complicated, but CPUE of the cluster six is a very low and this cluster separates the striped marlin distribution into the North and South Pacific. Furthermore, the range of the cluster 6 varies according to the season (Figure.7).
- The catch size information Japanese longline fishery changes in area and quarter, but the effect of year is small (Selectivity of Japanese longline fishery is not changed by annually).
- Focusing on WCNPO, clusters including age 0 fish appear in quarter 1 and 2 (Figure.8 , Figure.9).
- Based on these results, we propose 11 Japanese longline fleets for SS3.
- We also suggest two fleets for CPUE standardization. One is "Area 1 in quarter 1" where appears largest striped marlin (age three fish) and the other is the "Area 1 in quarter 3" that is the index for age one fish (Figure.9).

References

Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, and G. Grothendieck (2015). Package 'lme4'. *Convergence 12*(1).

Ijima, H. and M. Kanaiwa (2018). Pattern recognition of population dynamics for north pacific swordfish (xiphias gladius) : The operational data analysis of japanese longline fishery using the finite mixture model. *ISC/18/BILLWG-01/09*.

ISC (2015). Stock assessment update for striped marlin (kajika audax) in the western and central north pacific ocean through 2013.

Kinoshita, J., Y. Aoki, H. Ijima, and H. Kiyofuji (2018). Improvements in skipjack (katsuwonus pelamis) abundance index based on the fish size data from japanese pole-and-line logbook (1972–2017). *WCPFC-SC14-2018/ SA-WP-04 rev1*.

Langseth, B. (2016). Spatial and temporal patterns in striped marlin (kajikia audax) length in the hawaiian deep-set longline fishery. *ISC/16/BILLWG-01/08*.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent glass regression in r. *Journal of Statistical Software 11*.

Ochi, D., H. Ijima, J. Kinoshita, and H. Kiyofuji (2016). New fisheries definition from japanese longline north pacific albacore size data. *ISC/16/ALBWG-02/06*.

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

Waterhouse, L., D. B. Sampson, M. Maunder, and B. X. Semmens (2014). Using areas-as-fleets selectivity to model spatial fishing: asymptotic curves are unlikely under equilibrium conditions. *Fisheries Research 158*, 15–25.

Table 1: Candidate finite mixture models. "fixed" regression is not changed by throughout the cluster.

Modle	Cluster	GLM
m1	5X5 qtr	stm∼yr, fixed=∼gear, offset=effort, family="poisson", link="log" log(stmw)∼1, fixed=∼ yr+gear, family="gaussian", link="identity"
m2	5X5 qtr	stm∼yr+gear, offset=effort, family="poisson", link="log" log(stmw)∼yr+gear, family="gaussian", link="identity"
m2rev	5X5 qtr	stm∼yr+gear, offset=effort, family="poisson", link="log" log(stmw)∼gear, family="gaussian", link="identity"
m3	5X5	stm∼yr+gear+qtr, offset=effort, family="poisson", link="log" log(stmw)∼yr+gear+qtr, family="gaussian", link="identity"
m4	5X5 qtr	stm∼yr,fixed=∼gear, offset=effort, family="poisson", link="log" log(stmw)∼yr, fixed=∼gear, family="gaussian", link="identity"
m5	5X5 qtr	stm∼1, fixed =∼yr+gear, offset=effort, family="poisson", link="log" log(stmw)∼1, fixed=∼ yr+gear, family="gaussian", link="identity"

Table 2: Deviance table of selected model m2rev. The cluster gave a prior value of 2 to 12 and selected the finite mixture model with the smallest BIC.

Prior	Estimate	logLik	Deviance	Residual Deviance	AIC	BIC
2	2	−431464.8	862787.4	780619.9	863043.7	863477.6
3	3	−361196.7	703709.7	621542.2	722565.5	723220.2
4	4	−312812.4	603692.1	521524.6	625854.9	626730.3
5	5	−284960.3	566157.5	483990.0	570208.6	571304.8
6	6	−265727.8	529694.6	447527.1	531801.6	533118.6
7	7	−250968.1	538659.2	456491.7	502340.1	503877.9
8	8	−246823.9	506435.4	424267.9	494109.9	495868.4
9	8	−251021.0	495770.9	413603.4	502503.9	504262.4
10	9	−229463.0	476058.4	393890.9	459446.0	461425.3
11	7	−260140.7	478088.6	395921.1	520685.4	522223.2
12	9	−232601.7	481060.9	398893.4	465723.3	467702.6

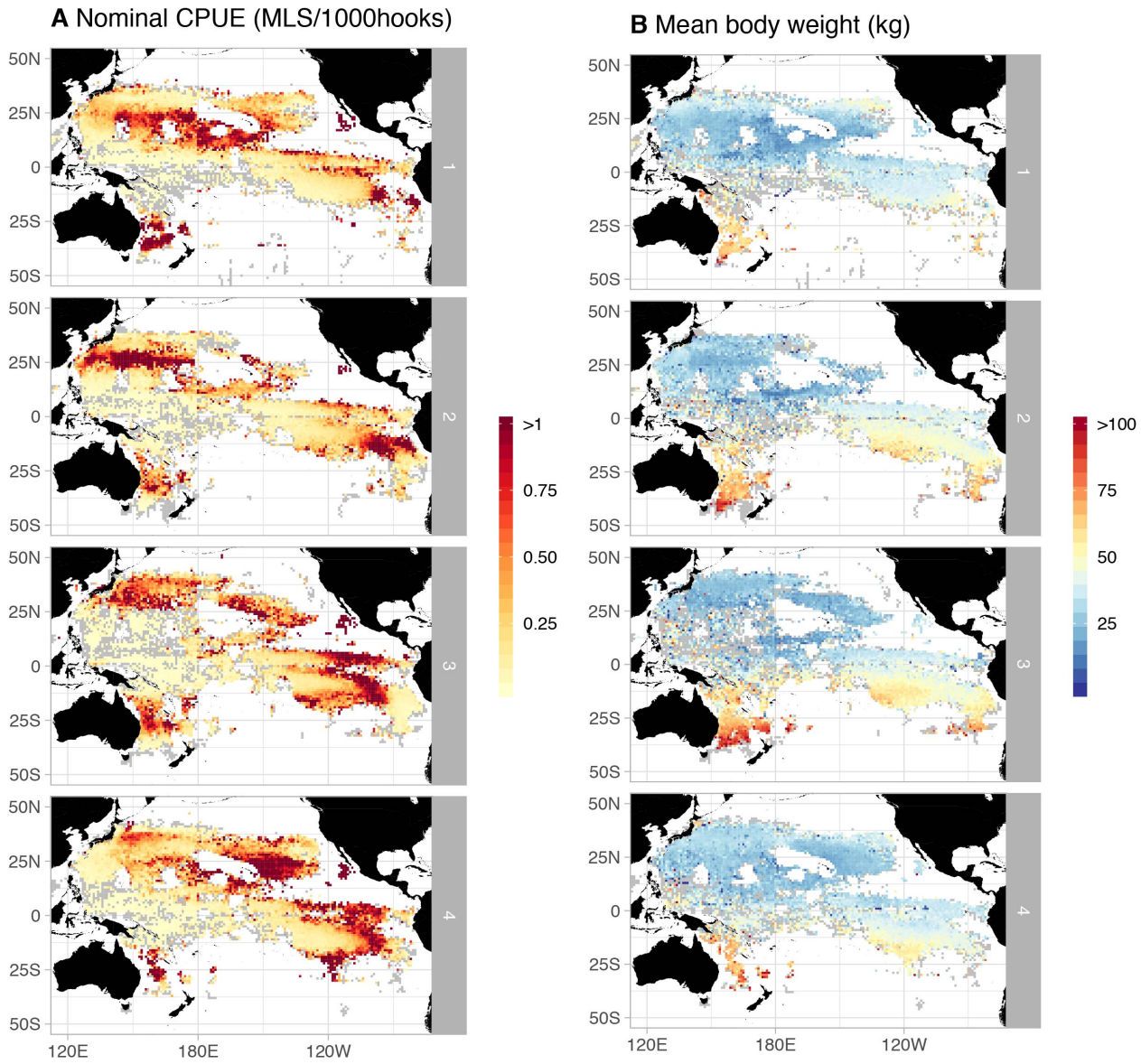

Figure 1: Quarterly catch pattern of Pacific striped marlin that was summarised by Japanese longline logbook data (1994-2017). The spatial resolution is 1 ° × 1 ° gridgrid. Mean body weight was calculated by the total catch amount divided by total catch number. 1 ° × 1 ° grid filled with gray indicate zero catch for Pacific striped marlin between 1994 and 2017.

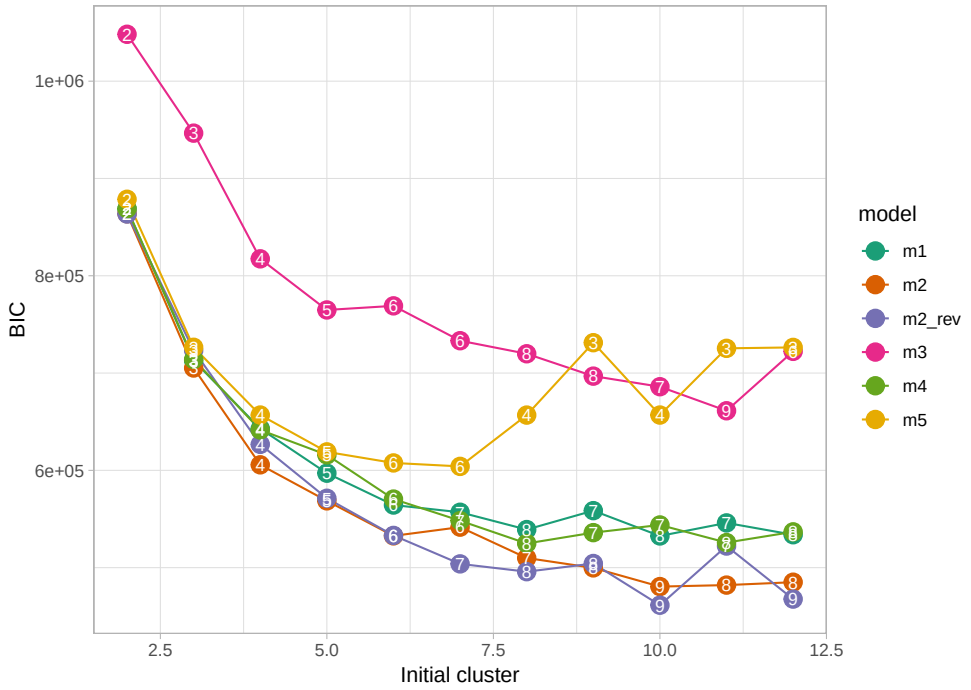

Figure 2: Bayesian Information Criterion (BIC) for each finite mixture models. Numbers in the circle denote the estimated number of clusters.

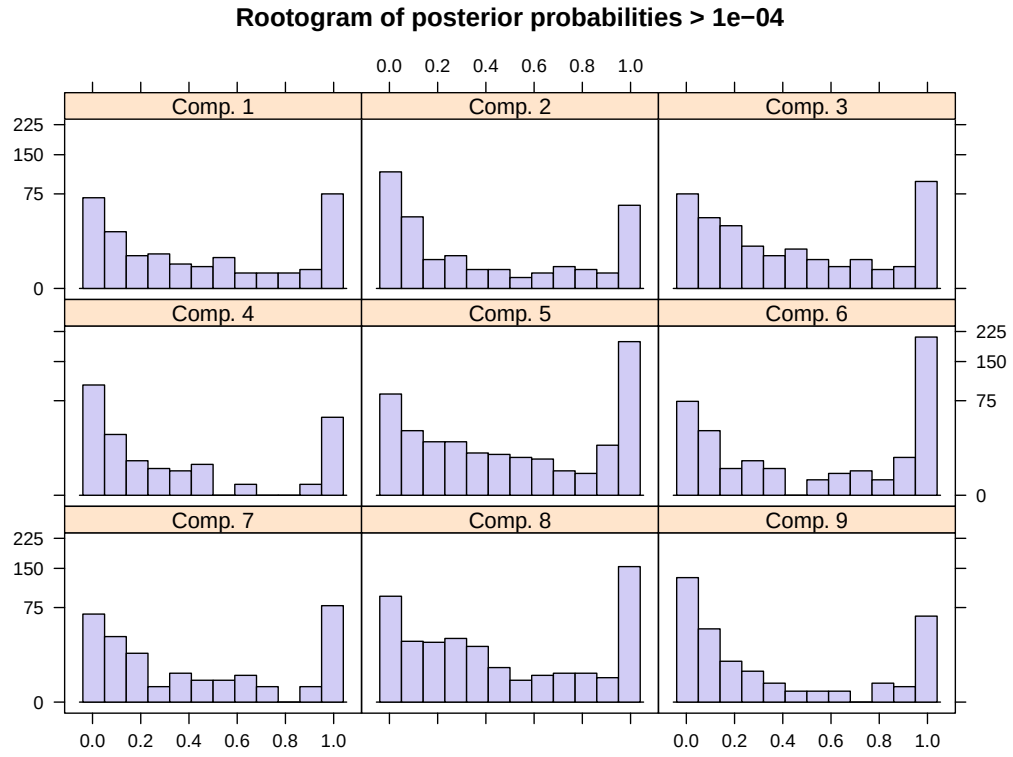

Figure 3: The rootogram of posterior probabilities. Posterior is summarized by larger than 0.0001.

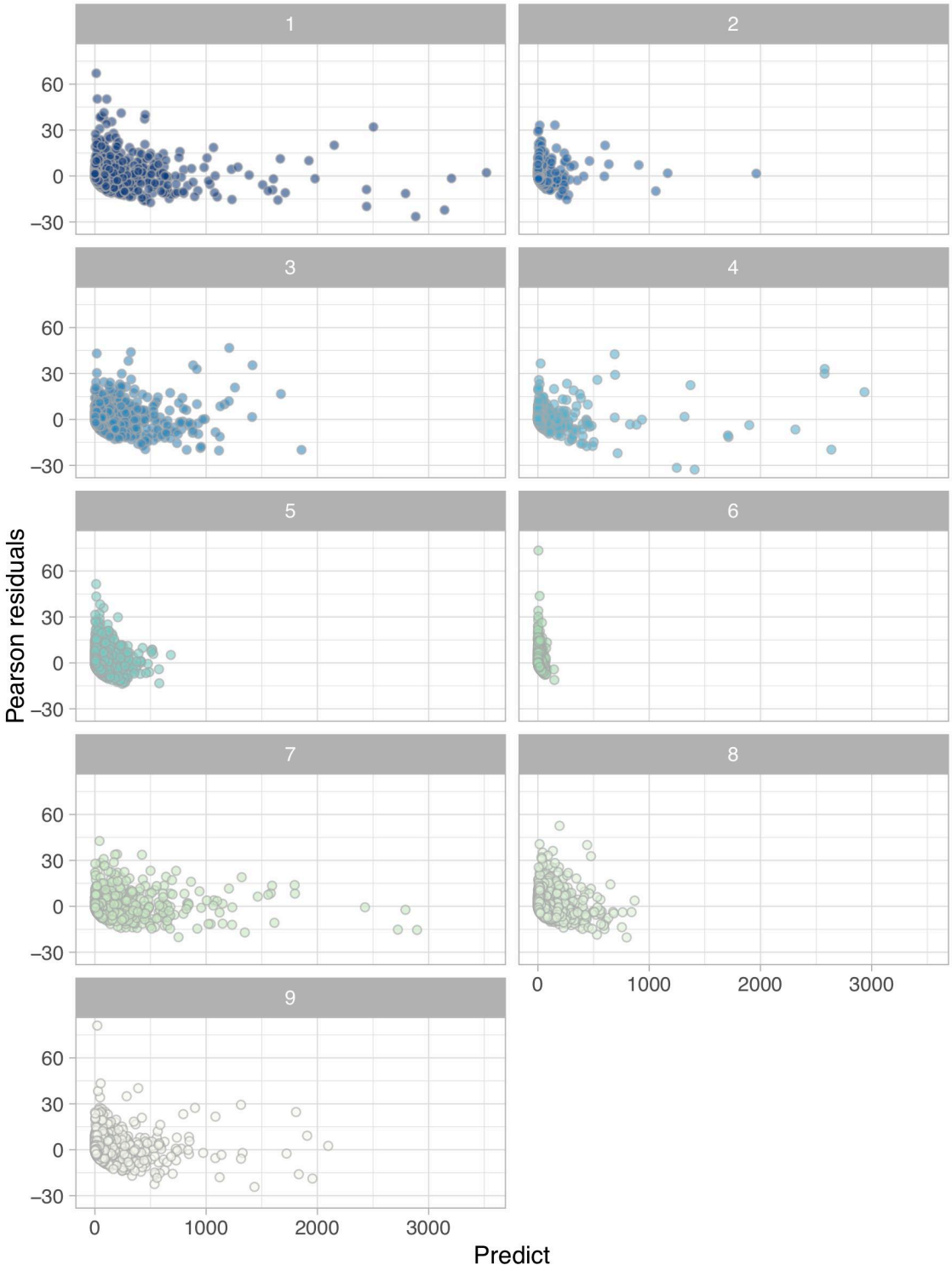

Figure 4: Pearson residuals of the CPUE GLM (m2rev). Pearson residuals aggregated by predicted nine clusters.

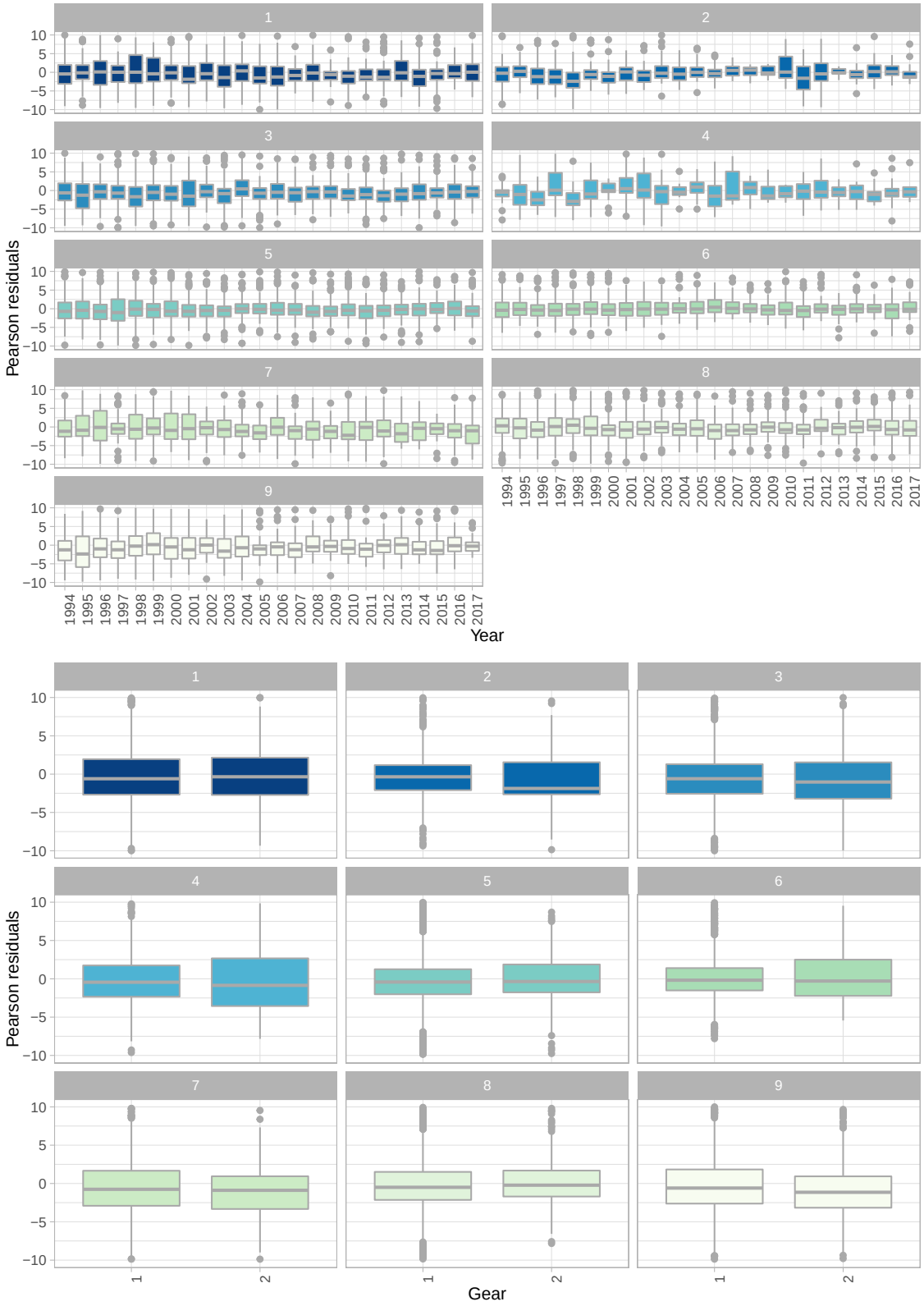

Figure 5: Pearson residuals of the CPUE GLM (m2rev). Top panel: Pearson residuals aggregated by year. Lower panel: Pearson residuals aggregated by gear.

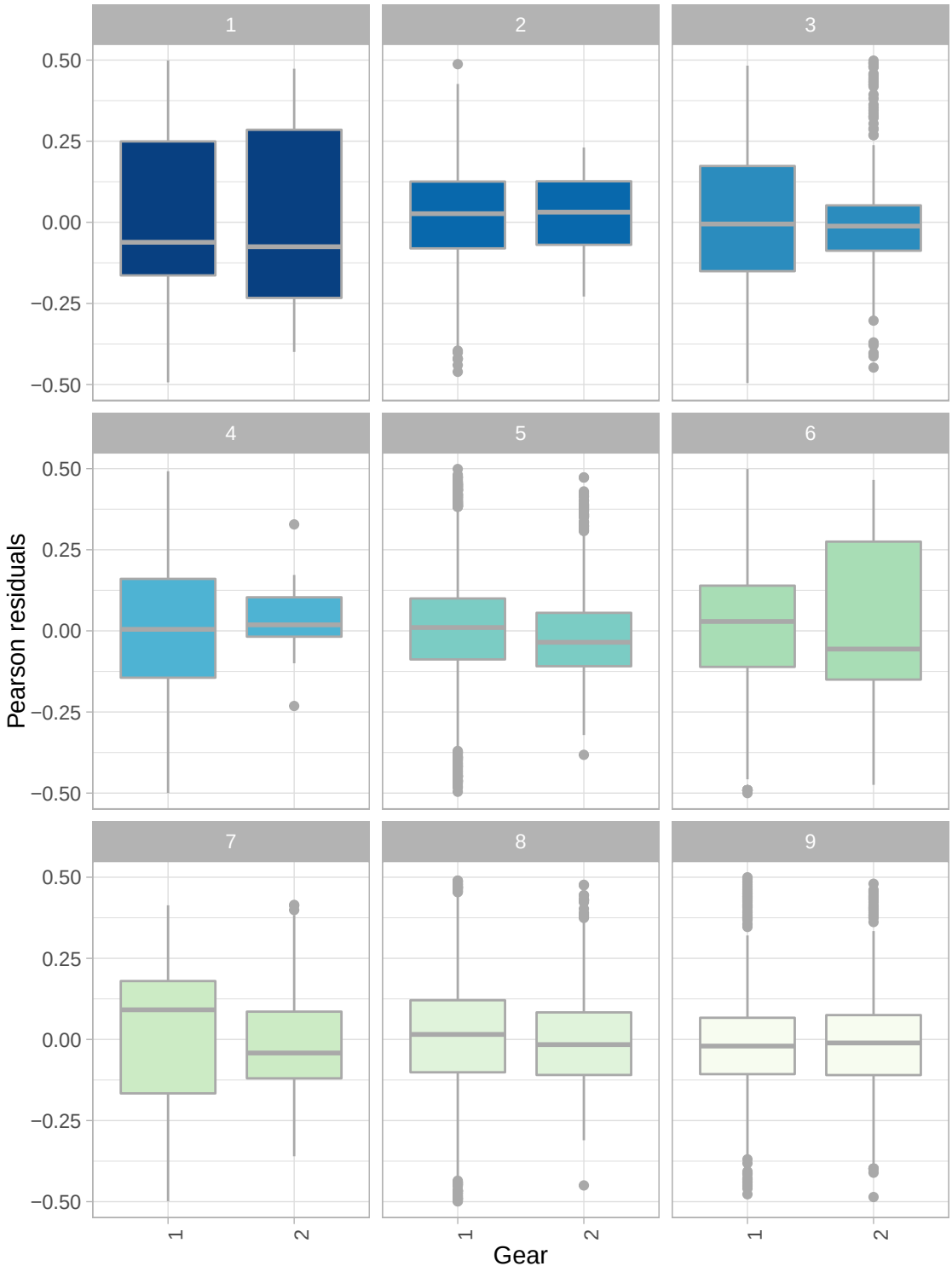

Figure 6: Pearson residuals of the body weight GLM (m2rev). We summarized the Pearson residuals by gear because the covariate of body weight GLM is only gear.

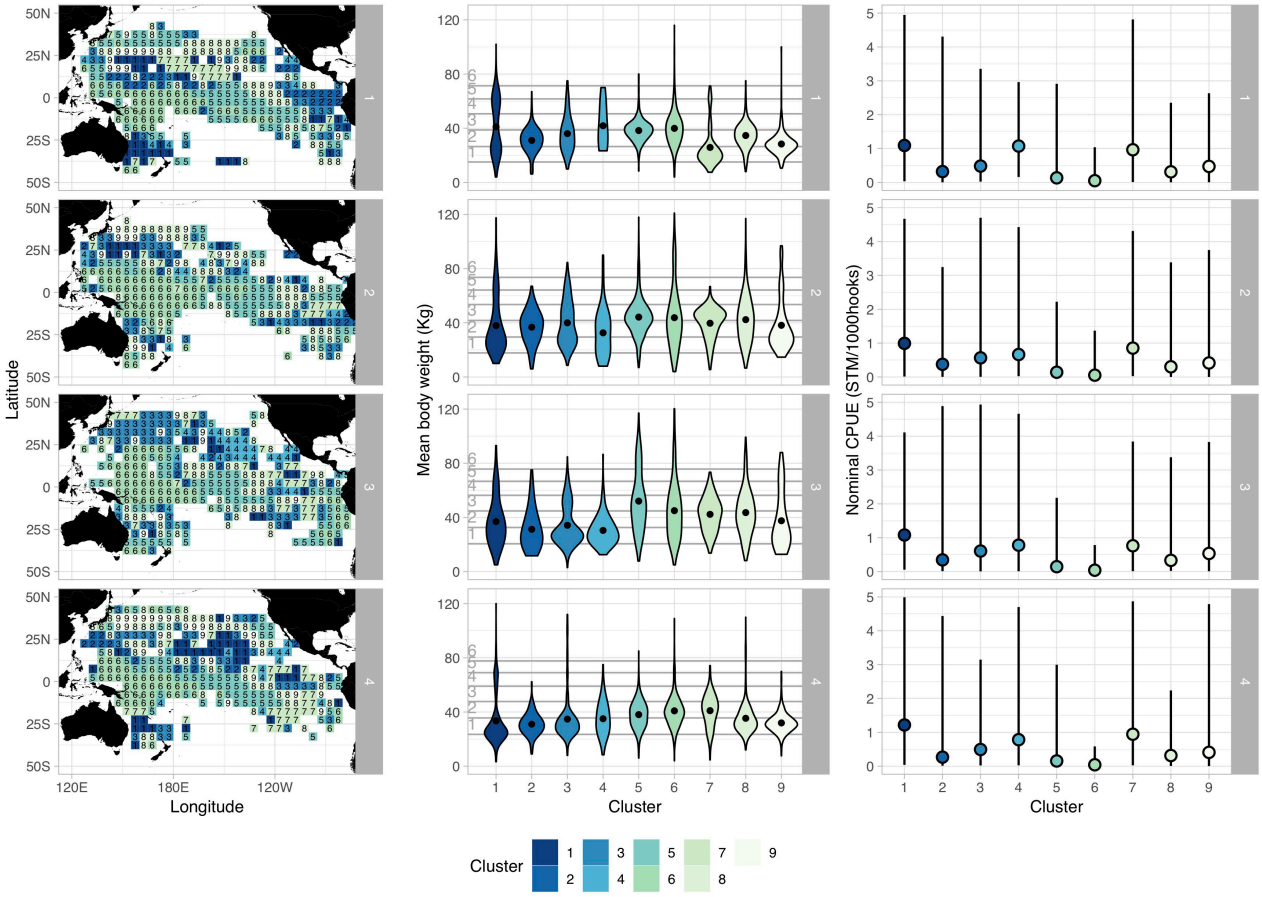

Figure 7: The distribution pattern of Pacific striped marlin. Finite mixture model estimates nine different clusters for each season. Left panel: Estimated cluster. Center panel: Violin plot of mean body weight. Mean body weight is calculated by 5 ° × 5 ° grid area. Right panel: Pattern of CPUE by each cluster. Circle denotes mean CPUE, error bar means max-minimum CPUE. All CPUEs are calculated by 5 ° × 5 ° grid area.

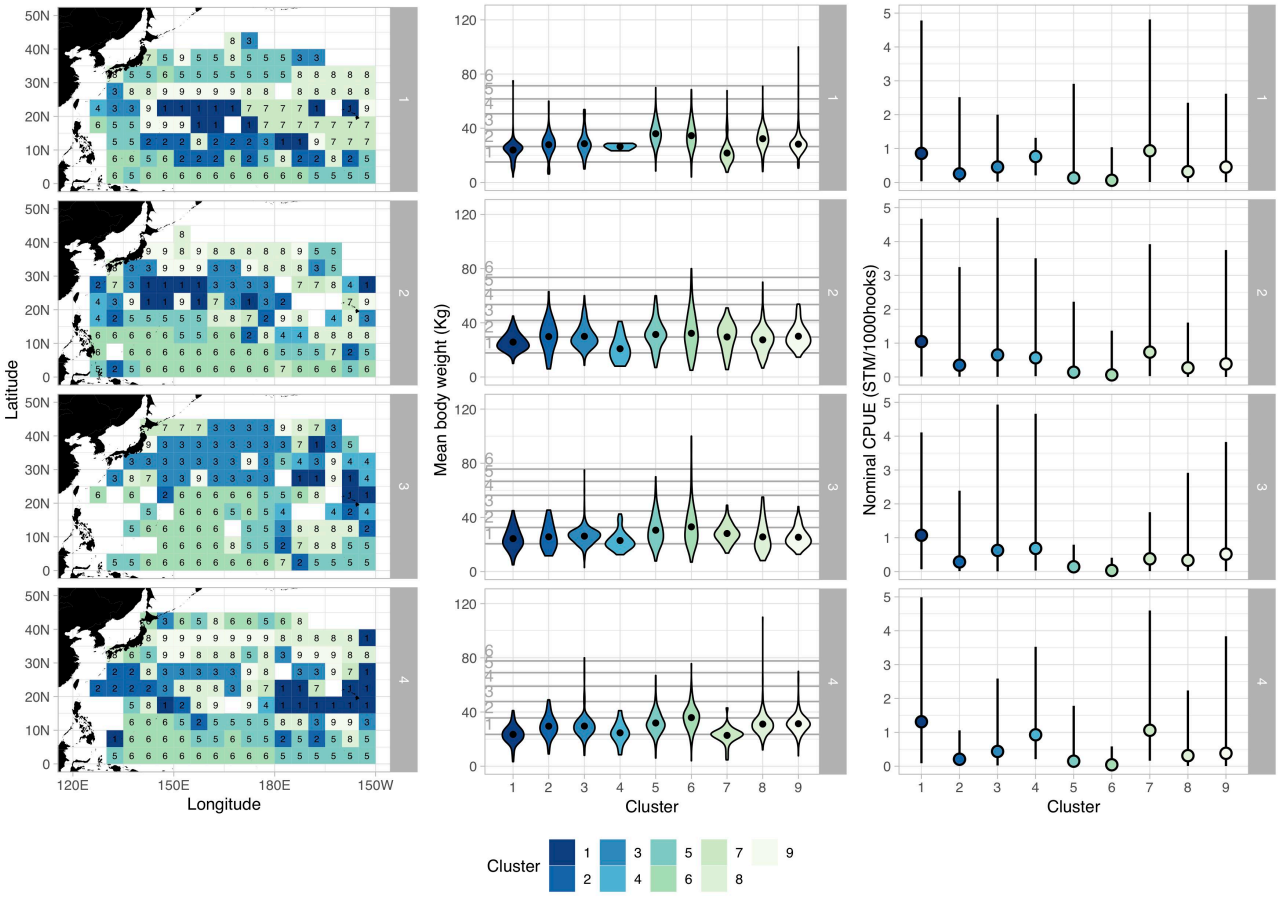

Figure 8: The distribution pattern of Pacific striped marlin focus on Western Central North Pacific Ocean. Left panel: Estimated cluster. Center panel: Violin plot of mean body weight by cluster. Mean body weight is calculated by 5 ° × 5 ° grid area. Right panel: Pattern of CPUE by each cluster. Circle denotes mean CPUE, error bar means max-minimum CPUE. All CPUEs are calculated by 5 ° × 5 ° grid area.

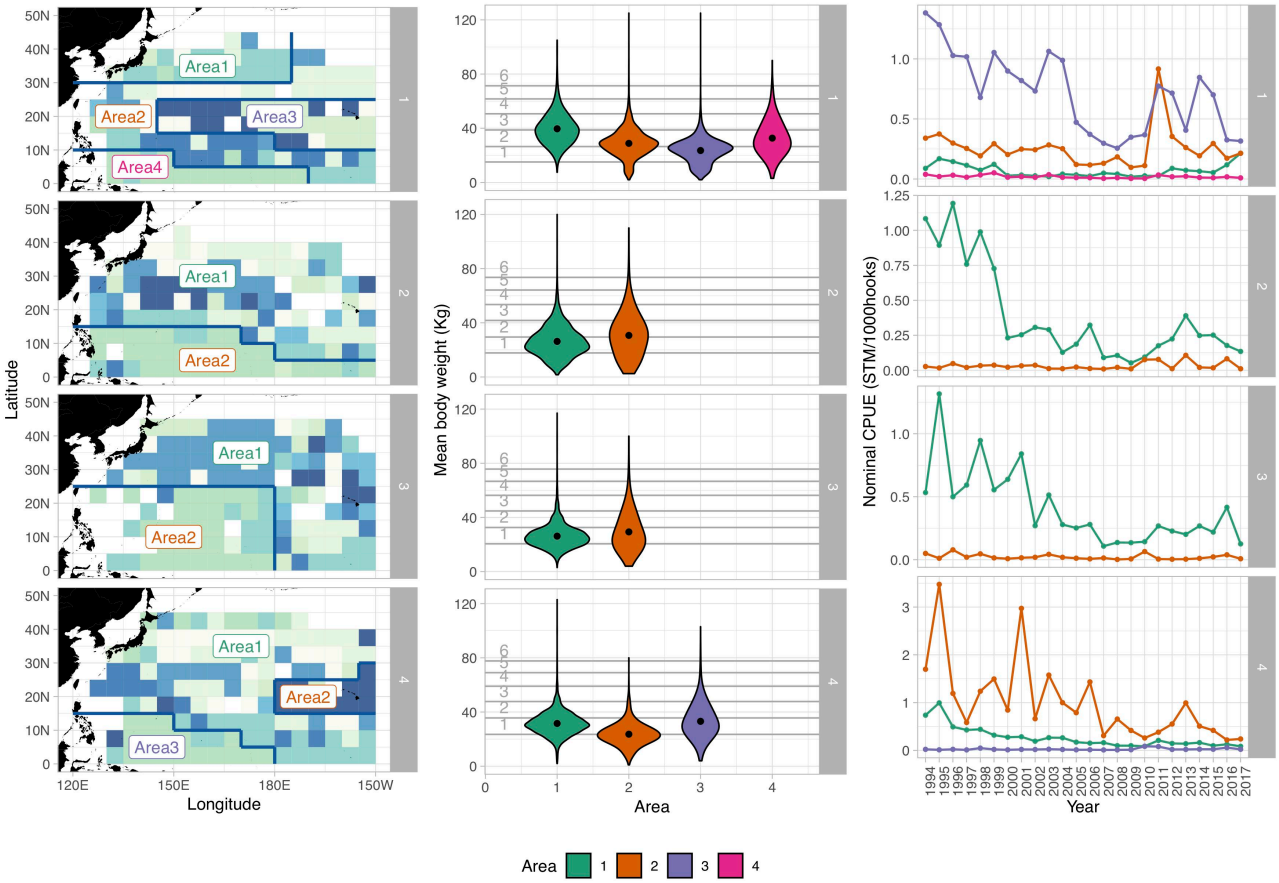

Figure 9: Japanese longline fleet definition for the stock synthesis 3. Considering the results of finite mixture model analysis, we defined 11 fleets. Left panel: The area-seasonal fleet definition for WCNPO stripe marlin longline fishery. Center panel: Violin plot of mean body weight by defined area. Mean body weight is calculated by $1\ ^{\circ} \times 1\ ^{\circ}$ grid area. Right panel: Trends of nominal CPUE by defined area.