Standardization of albacore CPUE by Japanese longline in the Indian Ocean which includes cluster analysis

Takayuki Matsumoto[1] and Simon Hoyle[2]

[1]*National Research Institute of Far Seas Fisheries (NRIFSF), Japan Fisheries Research and Education Agency, 5-7-1, Orido, Shimizu, Shizuoka, 424-8633, Japan*

[2]*IOTC consultant, 14 Champion Terrace, Nelson, New Zealand*

Abstract

Standardizations of Japanese longline CPUE for albacore in the Indian Ocean regions were conducted. The models incorporated fishing power based on vessel ID where available, and used cluster analysis to account for targeting. The variables year-quarter, vessel ID, latlong5 (five degree latitude-longitude block), cluster and number of hooks were used in the standardization. The numbers of clusters selected varied among regions, but in all cases were either 4 or 5. Dominant species differed by cluster. The effects of each covariate differed depending on region. The CPUEs trends were decreasing during the early period, and constant or increasing after that.

1. Introduction

Until 2016, national scientists have mainly standardized Japanese longline CPUE for albacore in the Indian Ocean using generalized linear models (GLM), with log normal errors and either operational or aggregated catch and effort data (e.g. Matsumoto and Uosaki 2011; Matsumoto et al. 2012, 2014, Matsumoto and Kitakado 2016). The standardizations have incorporated the effects of fishing season, area, fishing gear (number of hooks between floats and gear material) and an environmental factor (sea surface temperature). These may be termed ‘simple’ and ‘traditional’ methods.

In 2016, IOTC joint CPUE analysis (CPUE workshop) was conducted and ‘joint CPUEs’ were created for albacore, based on Japanese, Taiwanese and Korean longline operational data (Hoyle et al., 2016). These models account for fishing power based on vessel ID where available, and use cluster analysis to incorporate targeting. Joint CPUEs were considered to be more representative of status of the stocks and so were used for base models of stock assessment. At that time fleet-specific CPUE indices including Japanese longline CPUE were not prepared, so it was not possible to compare the joint and Japanese-only longline CPUE indices or CPUE indices among fleet based on the same method. In 2018 the joint CPUE analysis workshop was held and albacore CPUE indices for each fleet as well as joint CPUE were created (IOTC 2018). This document reports the standardization of albacore CPUE by Japanese longline conducted at this year’s joint CPUE analysis, using the same methods as that for joint CPUE.

2. Materials and methods

Data

Operational level (set by set) Japanese longline logbook data were used. The data were available for 1952-2017 (data for 2017 were preliminary), with the fields year, month and day of operation, location to 1° of latitude and longitude, vessel call sign, no. of hooks between floats (HBF), number of hooks per set, date of the start of the fishing cruise, logbook identifier, and catch in number of each species. Vessel call signs were available from 1979 onward and were used for the vessel identifier. The operations with hooks per set above 5000 and less than 200 were removed. Sets after 1975 with HBF missing or > 25 were removed. Sets before 1975 with missing HBF were allocated HBF of 5, according to standard practice with Japanese longline data.

Each set was allocated to albacore regions (Fig. 1). At the last (2016) IOTC albacore assessment, similar regions were used for longline joint CPUE but the was no southern limit for regions 3 and 4. This time, south of 40S was eliminated because this area was considered as fishing ground almost only for southern bluefin tuna.

Cluster analysis

We clustered the data using the approach applied by Hoyle et al. (2015). We removed all sets with no catch of any of the species, and then aggregated by vessel-month. Set level data contains variability in species composition due to the randomness of chance encounters between fishing gear and schools of fish. This variability leads to some misallocation of sets using different fishing strategies. Aggregating the data tends to reduce the variability, and therefore reduce misallocation of sets. For these analyses we aggregated the data by vessel-month, assuming that individual vessels tend to follow a consistent fishing strategy through time. One trade-off with aggregation in this way is that vessels may change their fishing strategy within a month, which will result in misallocation of sets. For the purposes of this paper we refer to aggregation by vessel-month as trip-level aggregation, although the time scale is (for distant water vessels) in most cases shorter than a fishing trip. In the data prior to 1979 vessel id was not available, but we were able to cluster them by vessel-month because the logbook id, available for the first time in the current data set, could be used to identify sets on the same vessel-trip.

We calculated proportional species composition by dividing the catch in numbers of each species by catch in numbers of all species in the vessel-month. Thus the species composition values of each vessel-month summed to 1, ensuring that large catches and small catches were given equivalent weight. The data were transformed by centering and scaling, so as to reduce the dominance of species with higher average catches. Centering was performed by subtracting the column (species) mean from each column, and scaling was performed by dividing the centered columns by their standard deviations.

We clustered the data using the hierarchical Ward hclust method, implemented with function hclust in R, option 'Ward.D', after generating a Euclidean dissimilarity structure with function 'dist'. This approach differs from the standard Ward D method which can be implemented by either taking the square of the dissimilarity matrix or using method 'ward.D2' (Murtagh & Legendre 2014). However in practice the method gives similar patterns of clusters to other methods, more reliably than ward.D2 (Hoyle et al 2015).

Data were also clustered using the kmeans method, which minimises the sum of squares from points to the cluster centres, using the algorithm of Hartigan and Wong (1979). It was implemented using function kmeans in the R stats package (R Core Team 2014).

Selecting the number of groups

We used several subjective approaches to select the appropriate number of clusters. In most cases the approaches suggested the same or similar numbers of groups. First, we applied hclust to transformed trip-level data and examined the hierarchical trees, subjectively estimating the number of distinct branches. Second, we ran kmeans analyses on untransformed trip-level data with number of groups k ranging from 2 to 25, and plotted the deviance against k. The optimal group number was the lowest value of k after which the rate of decline of deviance became slower and smoother. Third, following Winker et al (2014) we applied the nScree() function from the R nFactors package (Raiche & Magis 2010), which uses various approaches (Scree test, Kaiser rule, parallel analysis, optimal coordinates, acceleration factor) to estimate the number of components to retain in an exploratory PCA. Where there was uncertainty about the number of clusters, we selected the option with more clusters.

We plotted the hclust clusters to explore the relationships between them and the species composition and other variables, such as HBF, number of hooks, year, and set location. Plots included boxplots of a) proportion of each species in the catch, by cluster; b) the distributions of variables by cluster; and c) maps of the spatial distribution of clusters, one map for each cluster.

In some analyses clusters that caught very few of the species of interest were omitted, because they provide little relevant information and may cause analysis problems due to large numbers of zeroes, and memory problems due to large sample sizes. Cluster selection was based on review and discussion of the plots of covariates and species compositions by cluster. Also, in some analyses only clusters in which target species was dominant were selected to see the difference from the results in which all or almost all the clusters were used. Analyses were run both with and without these clusters.

For standardization of each region, data were selected for vessels that had fished for at least N1 quarters in that region. The standard level of N1 was 8 quarters in the equatorial regions and 2 quarters in the southern regions. Subsequently, vessels, 5° cells, and year-quarters were included if they had at least 100 sets. For analyses of the 1952-1979 period this criterion was reduced to 50 sets, to increase the size of the dataset. For datasets with more than 60,000 sets the number of sets in each stratum (5° square * year-quarter) was limited by randomly selecting 60 sets without replacement from strata with more than this number of sets. Testing suggested that this approach did not cause bias, and the effects on trends of random variation were reduced to very low levels at 30 sets per stratum (Hoyle & Okamoto 2011), suggesting that 60 sets was more than adequate.

CPUE standardization, and fleet efficiency analyses

CPUE standardization methods generally followed the approaches used by Hoyle et al. (2015). The operational data were standardized using generalized linear models in R.

GLM (generalized linear models) that assumed a lognormal and delta lognormal distribution was conducted, and in this report mostly the methods and results for lognormal distributions are shown with partly for delta lognormal distribution. In this approach the response variable log (CPUE+k) was used, and a Normal distribution assumed. The constant k, added to allow for modelling sets with zero catches of the species of interest, was 10% of the mean CPUE for all sets. CPUE was defined at the set level as catch in number divided by hooks set. The following models were used:

Lognormal

$\ln(CPUEs{+}k){\sim}\ yrqtr{+}vessid{+}latlong5{+}cluster{+}f(hooks){+}\epsilon$

Delta lognormal

$(CPUE{=}0){\sim}\ yrqtr{+}vessid{+}latlong5{+}f(hooks){+}\ cluster\ {+}\epsilon$

$\log(CPUE){\sim}\ yrqtr{+}vessid{+}latlong5{+}f(hooks){+}\ cluster\ {+}\epsilon$, for nonzero sets

where $yrqtr$: year and quarter; $vessid$: effect of vessel ID; $latlong5$: effect of five degree latitude and longitude; *cluster*: effect of cluster; $f(hooks)$: function of number of hooks modelled with a cubic spline; *g(HBF)*: function of the number of hooks between floats modelled with a cubic spline; ϵ: error term.

Data periods

Vessel identity information was only available from 1979, so could not be applied uniformly across all years. The discontinuity in 1979 could be addressed in several different ways. We therefore analyzed the data in several ways so as to provide the assessment scientists with appropriate data. For each of the approaches above, four analyses were carried out as shown below.

Analysis	Years	Vessel effects
1	1952-1979	No
2	1979-end	Yes
3	1952- end	No
4	1952- end	Yes

It is possible to standardize the time series with vessel effects by assigning an identical dummy value to all vessels without vessel identity information. This was done for analysis 3). However using a dummy value introduces several problems. First, not all vessels begin to report their callsign at once in 1979, and those that do are self-selected and not randomly selected from the vessel population. Therefore it cannot be assumed that fishing power remains constant after 1979 for the dummy vessel id, so the transition in 1979 may introduce a discontinuity into the time series. The discontinuity can be limited in scope by restricting the overlap between dummy and real vessel IDs to one year – 1979 – and removing sets with missing vessel IDs after this time. Secondly, residuals may be more variable before 1979, without a true vessel ID in the model, which can introduce bias into the standardization.

One approach for addressing the discontinuity in analysis 3) is to adjust the time period 1952-1978 so that the relative averages in 1978 and 1979 are the same as they are in analysis 4), without vessel effects. However we considered that a better approach may be to estimate two time series 1952-1978 without vessel effects, and a second time series 1979-2017 with vessel effects (omitting all sets without vessel IDs). These are analyses 1) and 2) above. Subsequently the analyst can use them as desired, for example concatenating them after adjusting the averages so that the estimates for 1979 are the same.

Indices of abundance

Indices of abundance were obtained by applying the R function predict.glm to model objects. Binomial time effects were obtained by generating time effects from the glm and adjusting them so that their mean was the proportion of positive sets across the whole dataset. The main aim with this approach is to obtain a CPUE that varies appropriately, since variability for a binomial is greater when the mean is at 0.5 than at 0.02 or 0.98, and the multiplicative effect of the variability is greater when the mean is lower. The outcomes were normalised and reported as relative CPUE with mean of 1.

Uncertainty estimates were provided by applying the R function predict.glm with type = ”terms” and se.fit=TRUE, and taking the standard error of the year-quarter effect. For the delta lognormal models we used only the uncertainty in the positive component. Uncertainty estimates from standardizing commercial logbook data are in general biased low and often ignored by assessment scientists, since they assume independence and ignore autocorrelation associated with (for example) consecutive sets by the same vessels in the same areas. There may be a very large mismatch between the observation error in CPUE indices and the process error in the indices that is estimated in the assessment. This is particularly true for distant water longline CPUE, where very large sample sizes generate small observation errors.

Residual distributions and Q-Q plots were produced for all but the binomial analyses. For the lognormal positive analyses that included cluster in the model, median residuals were plotted by cluster. For all lognormal positive analyses, residuals by year-quarter were plotted by flag; median residuals by year-quarter were plotted by flag; and median residuals by 5° cell were mapped onto a contour plot for each flag.

3. Results and discussion

The aim of the cluster analysis was first to identify separate fishing strategies in the data for each species, regional structure, fleet, and region, and so to better understand the fishing practices; and second to assign each unit of fishing effort to a particular fishing strategy, so that the clusters could be used in standardization.

Species compositions were plotted by cluster for each region and fleet, as were the relative distributions of covariates (**Fig. 2** and **Fig. 3**, respectively). Dominant species differed depending on clusters. Clusters with low levels of the target species were excluded from standardization datasets. Numbers of clusters were 4 or 5.

Fig. 4 shows density distribution of each cluster for albacore region. Usually the distribution differed by the clusters.

Fig. 5 shows the effect of each covariate (for lognormal model). Change in vessel effect differed depending on regions, and increased after 1990 in region 4. Difference by 5 degree block is observed. As for the effect of the cluster, albacore dominant cluster had highest effect.

Fig. 6 shows the trend of standardized CPUE, without and with vessel effects. The trend differs among regions, but CPUE usually shows decreasing trend in the early period. After that, CPUE is comparatively stable, but is increasing trend after 1990s in the region 4.

Fig. 7 shows distribution of standardized residuals and QQ plots for lognormal model.

4. References

Hartigan, J. A. and M. A. Wong. 1979. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society.Series C (Applied Statistics) 28(1): 100-108.

Hoyle SD, Okamoto H 2011. Analyses of Japanese longline operational catch and effort for bigeye and yellowfin tuna in the WCPO, WCPFC-SC7-SA-IP-01. Western and Central Pacific Fisheries Commission, 9th Scientific Committee. Pohnpei, Federated States of Micronesia.

Hoyle SD, Okamoto H, Yeh Y-m, Kim ZG, Lee SI, Sharma R 2015. IOTC–CPUEWS02 2015: Report of the 2nd CPUE Workshop on Longline Fisheries, 30 April – 2 May 2015. 126 p.

Hoyle, S., Chang, Y., Yeh, Y., Satoh, K., Matsumoto, T., Kim, D. N. and Lee, S. 2016. Collaborative study of albacore tuna CPUE from multiple Indian Ocean longline fleets. IOTC–2016–WPTmT06–19.

IOTC (2018) Report of the Fifth IOTC CPUE Workshop on Longline Fisheries. IOTC-2018-WPM09-INF05. 27p.

Matsumoto, T., T. Kitakado and H. Okamoto. 2012. Standardization of albacore CPUE by Japanese longline fishery in the Indian Ocean. IOTC–2012–WPTmT04–10. 16pp.

Matsumoto, T., T. Kitakado and T. Nishida. 2014. Standardization of albacore CPUE by Japanese longline fishery in the Indian Ocean. IOTC–2014–WPTmT05–18. 20pp.

Matsumoto, T. and K. Uosaki. 2011. Standardization of albacore CPUE by Japanese longline fishery in the Indian Ocean. IOTC–2011–WPTmT03–15. 10pp.

Matsumoto, T. and Kitakado, T. 2016. Standardization of albacore CPUE by Japanese longline fishery in the Indian Ocean. IOTC–2016–WPTmT06–15. pp 22.

Murtagh F, Legendre P 2014. Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? Journal of Classification 31(3): 274-295.

Raiche, G. and D. Magis. 2010. "nFactors: Parallel analysis and non graphical solutions to the Cattell Scree Test." R package version 2(3).

Winker H, Kerwath SE, Attwood CG 2014. Proof of concept for a novel procedure to standardize multispecies catch and effort data. Fisheries Research 155: 149-159.

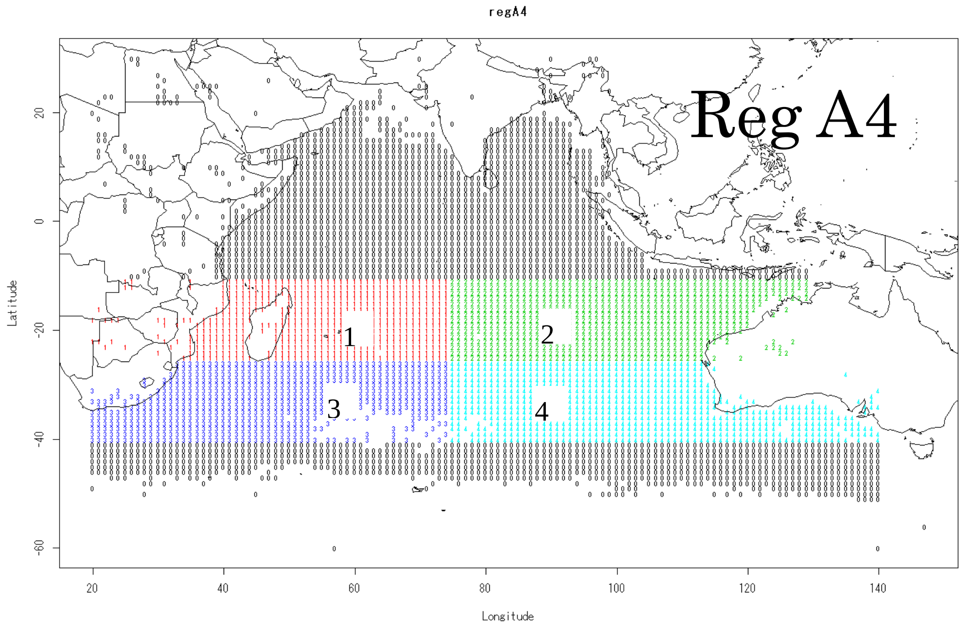


Fig. 1. Maps of the regional structures used to estimate albacore CPUE indices, regA4.

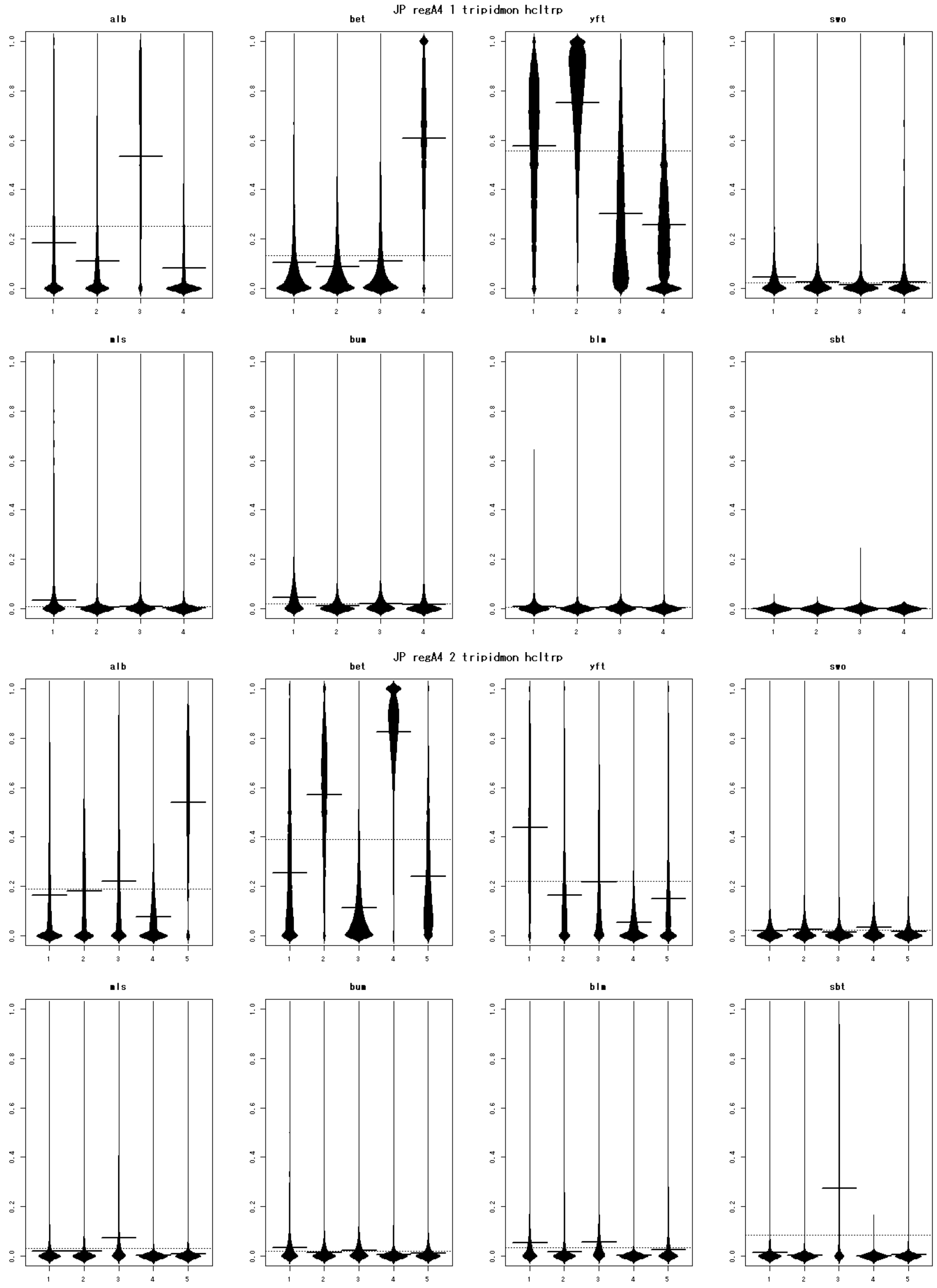


Fig. 2. Beanplots for albacore region showing species composition by cluster. The horizontal bars indicate the medians.

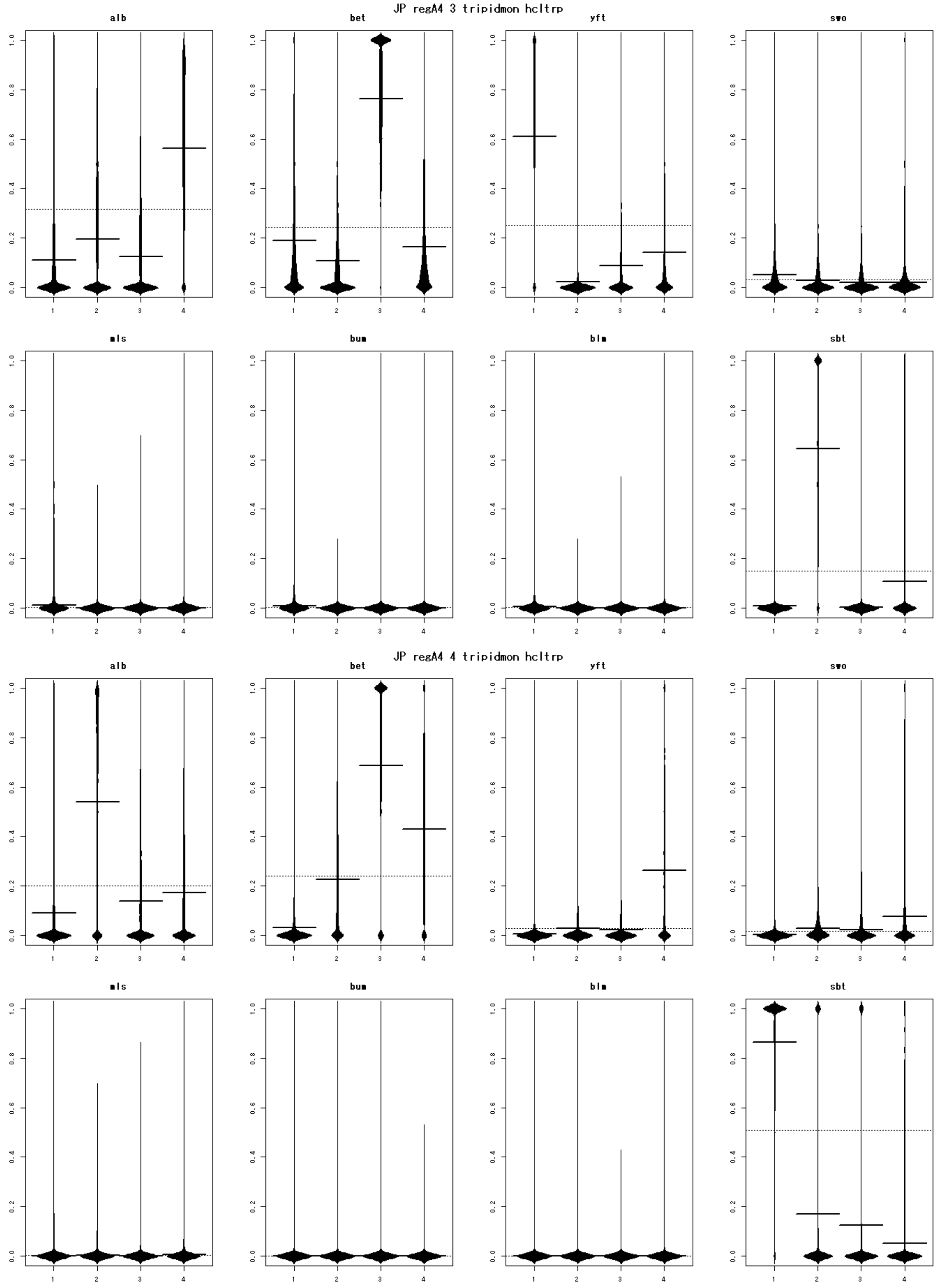


Fig. 2. Beanplots for albacore region showing species composition by cluster. The horizontal bars indicate the medians. (continued)

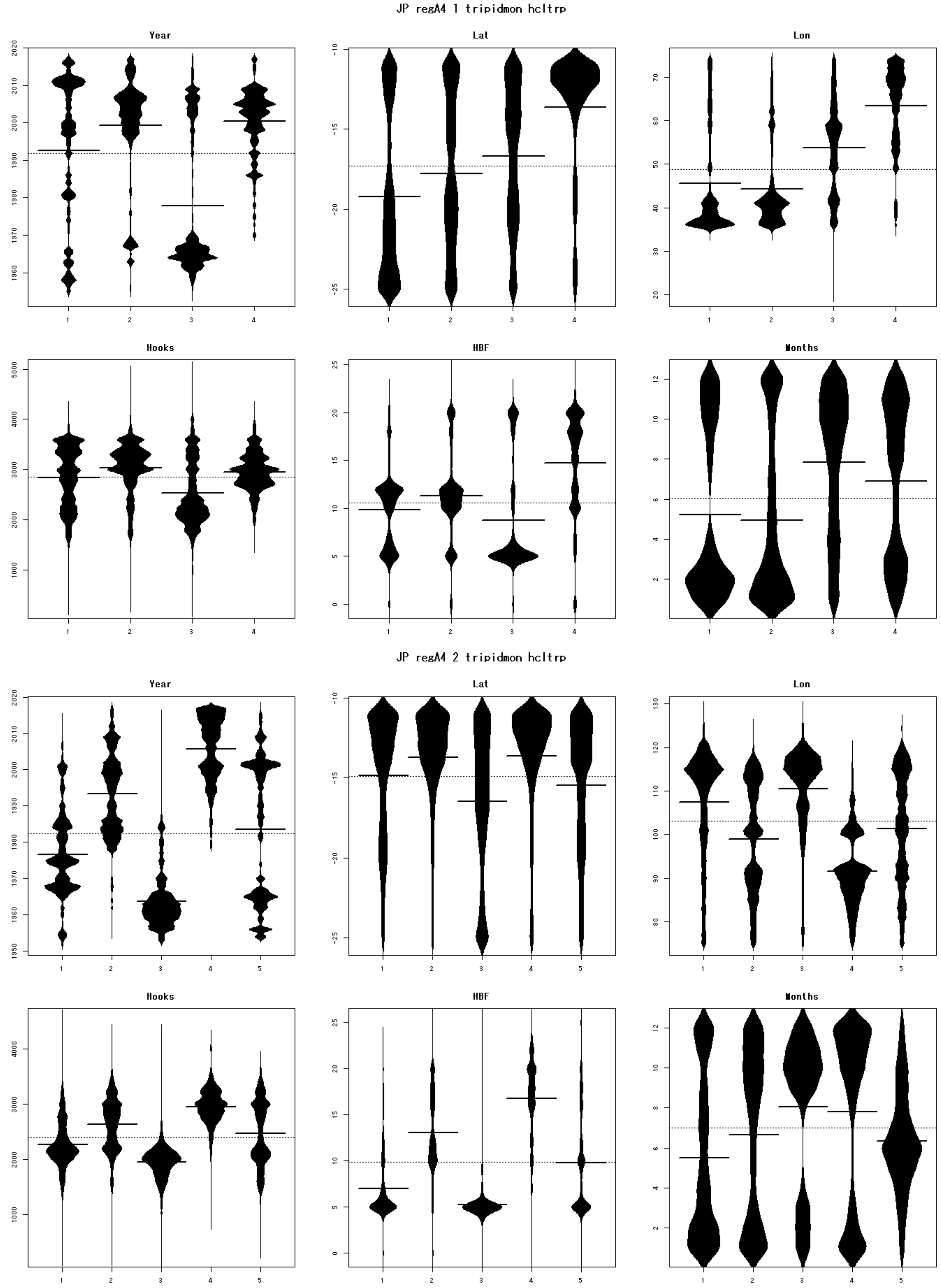


Fig. 3. Beanplots for albacore region showing number of sets versus covariate by cluster. The horizontal bars indicate the medians.;;

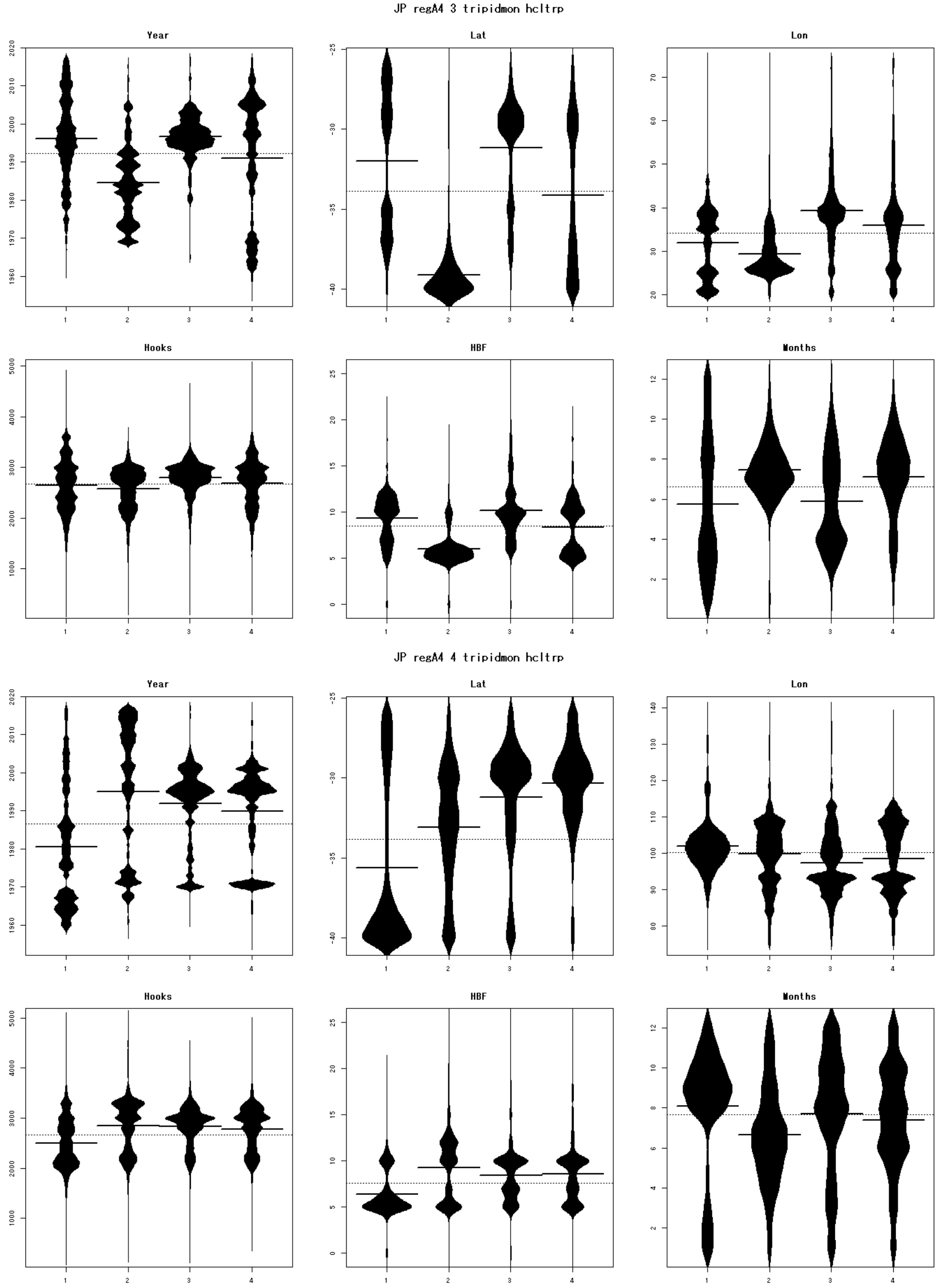


Fig. 3. Beanplots for albacore region showing number of sets versus covariate by cluster. The horizontal bars indicate the medians. (continued)

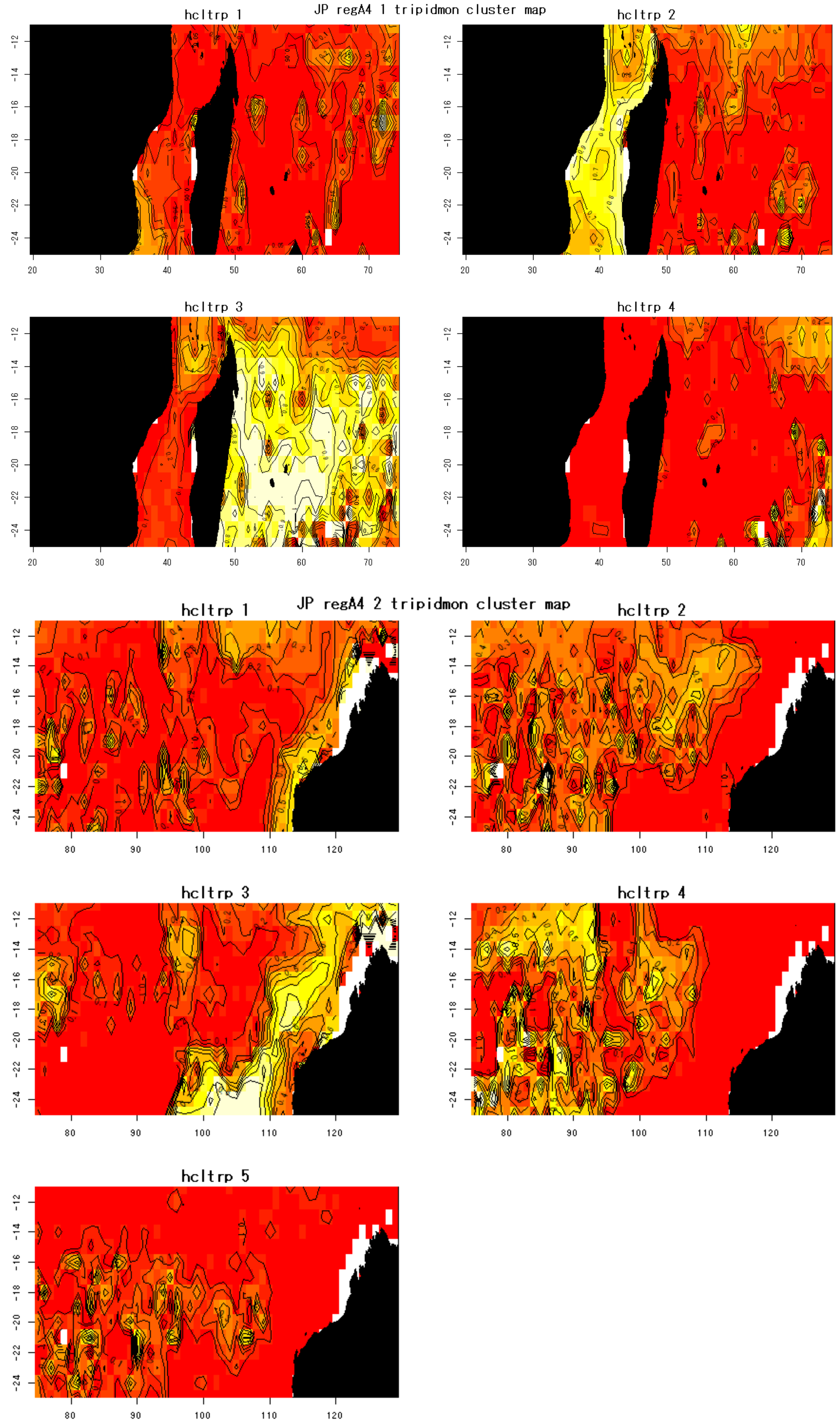


Fig. 4. Distribution for each cluster in albacore region showing the density of cluster. Yellow color is higher density.

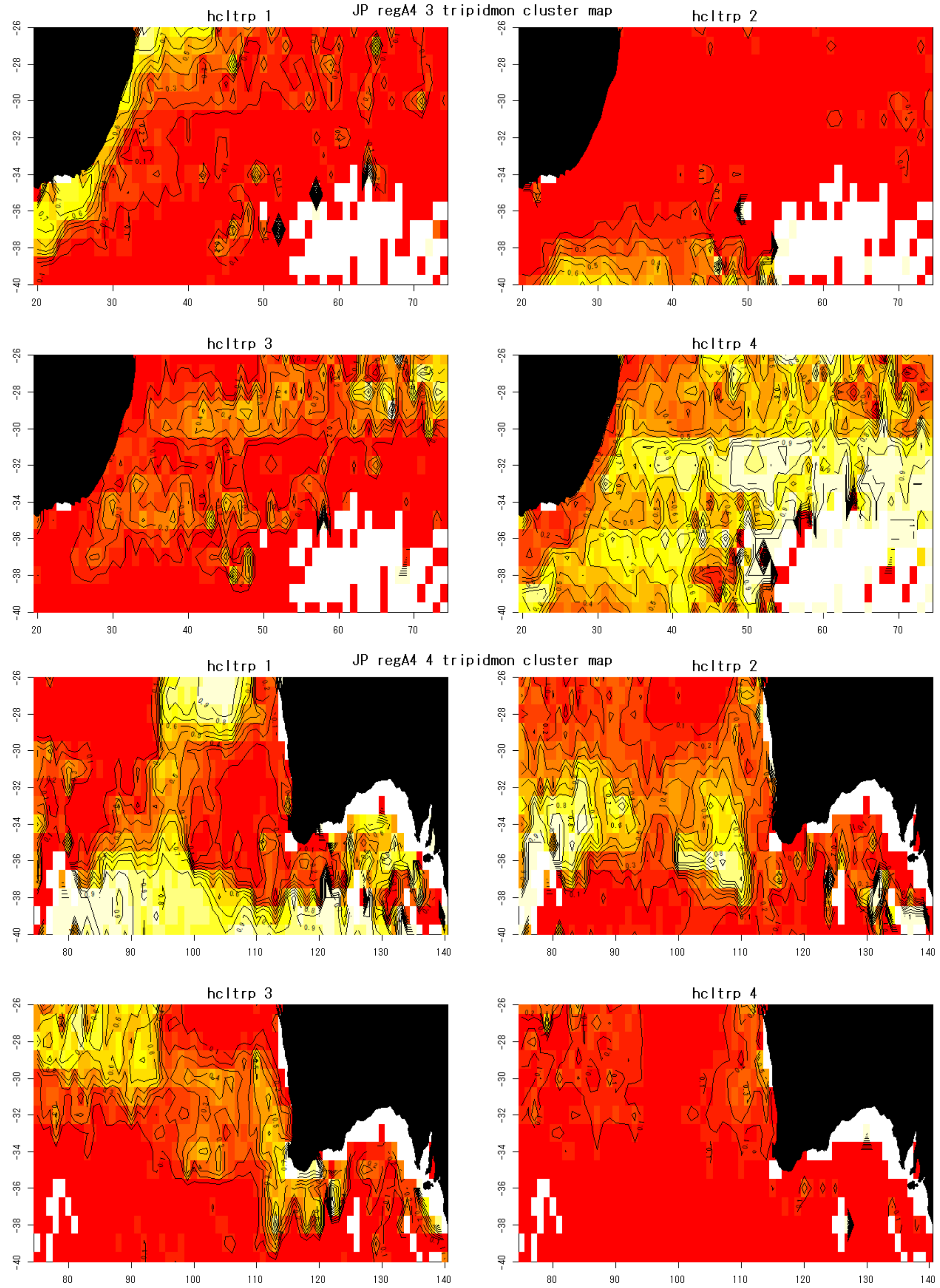
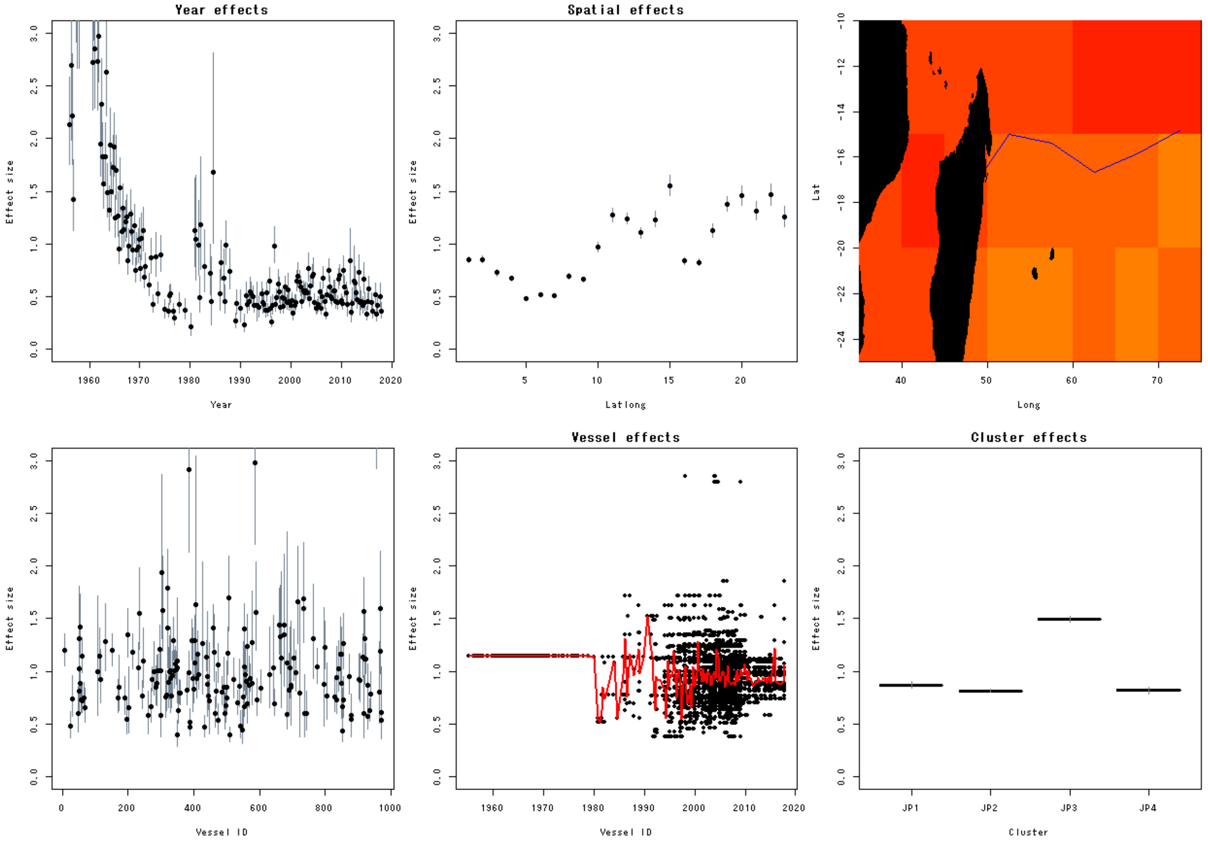


Fig. 4. Distribution for each cluster in albacore region showing the density of cluster. Yellow color is higher density. (continued)

Region 1



Region 2

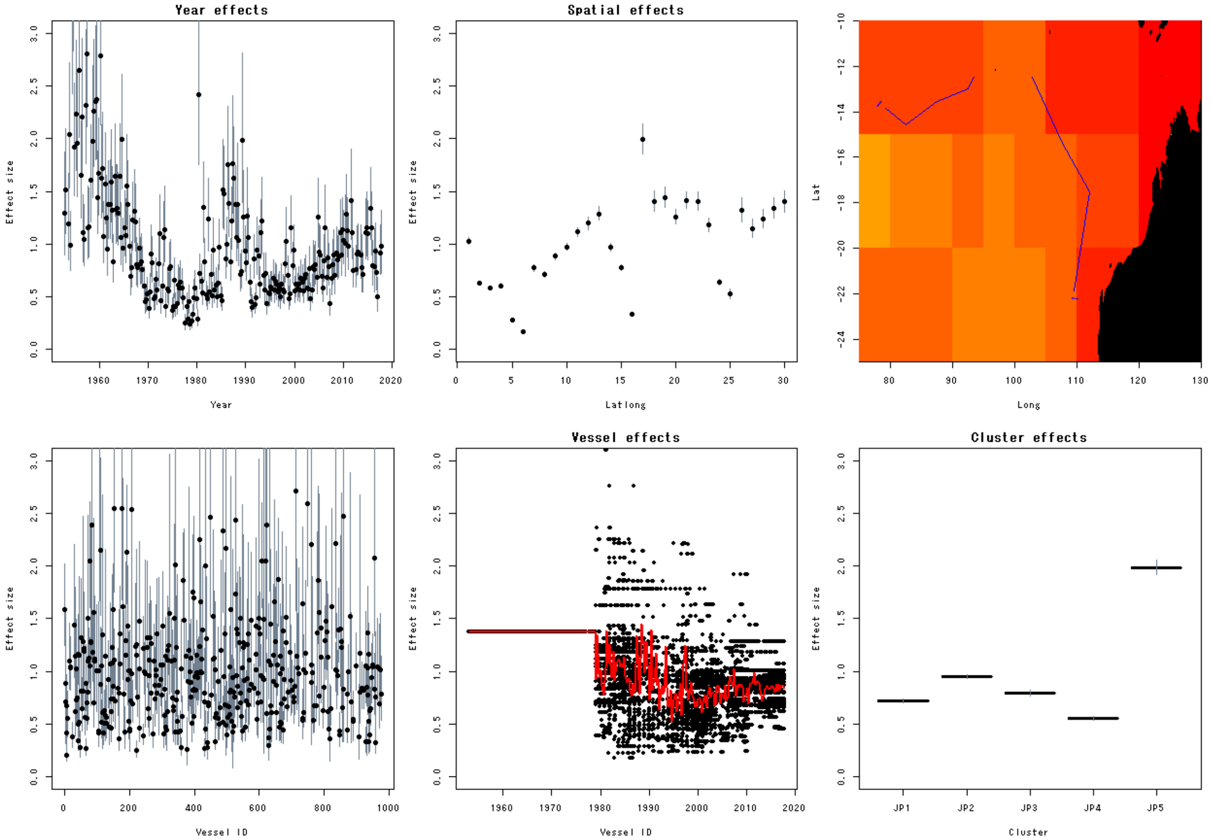
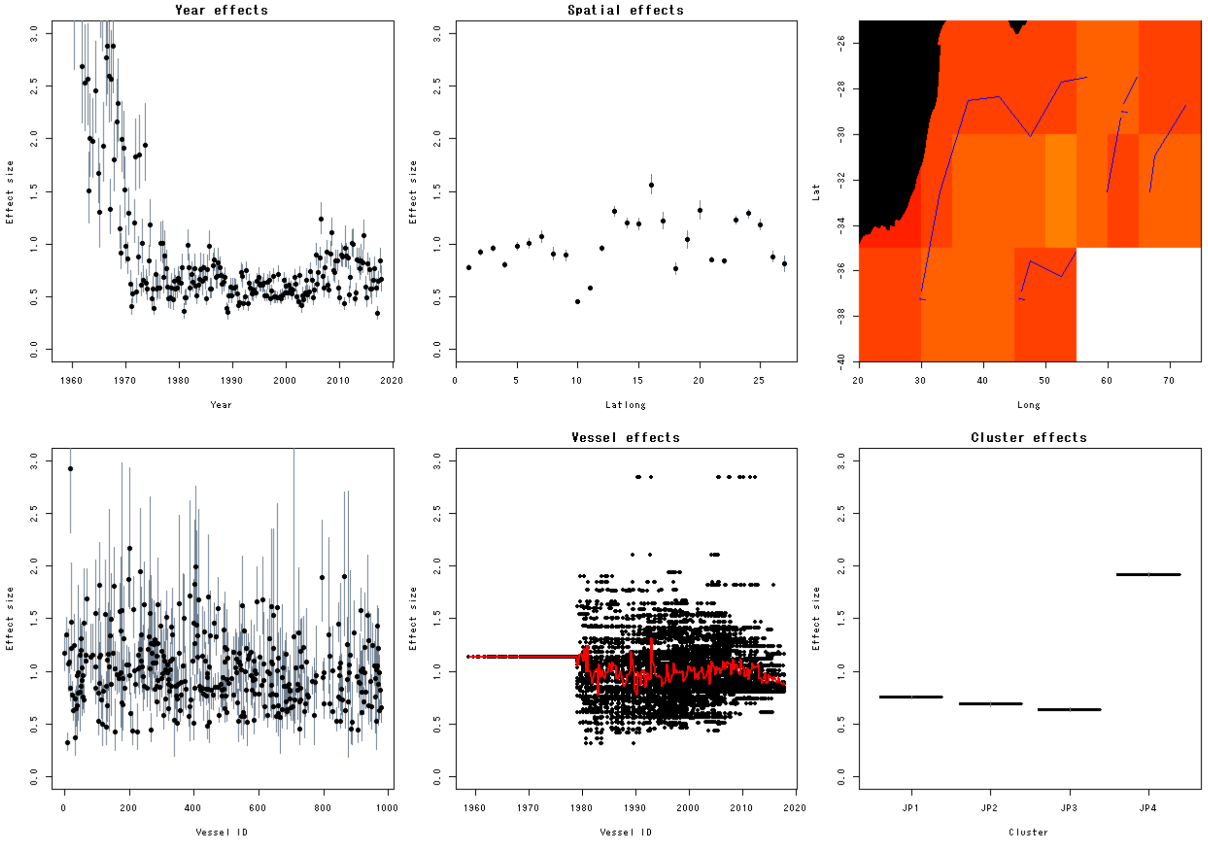


Fig. 5. Effect of each covariate for bigeye region (lognormal model).

Region 3



Region 4

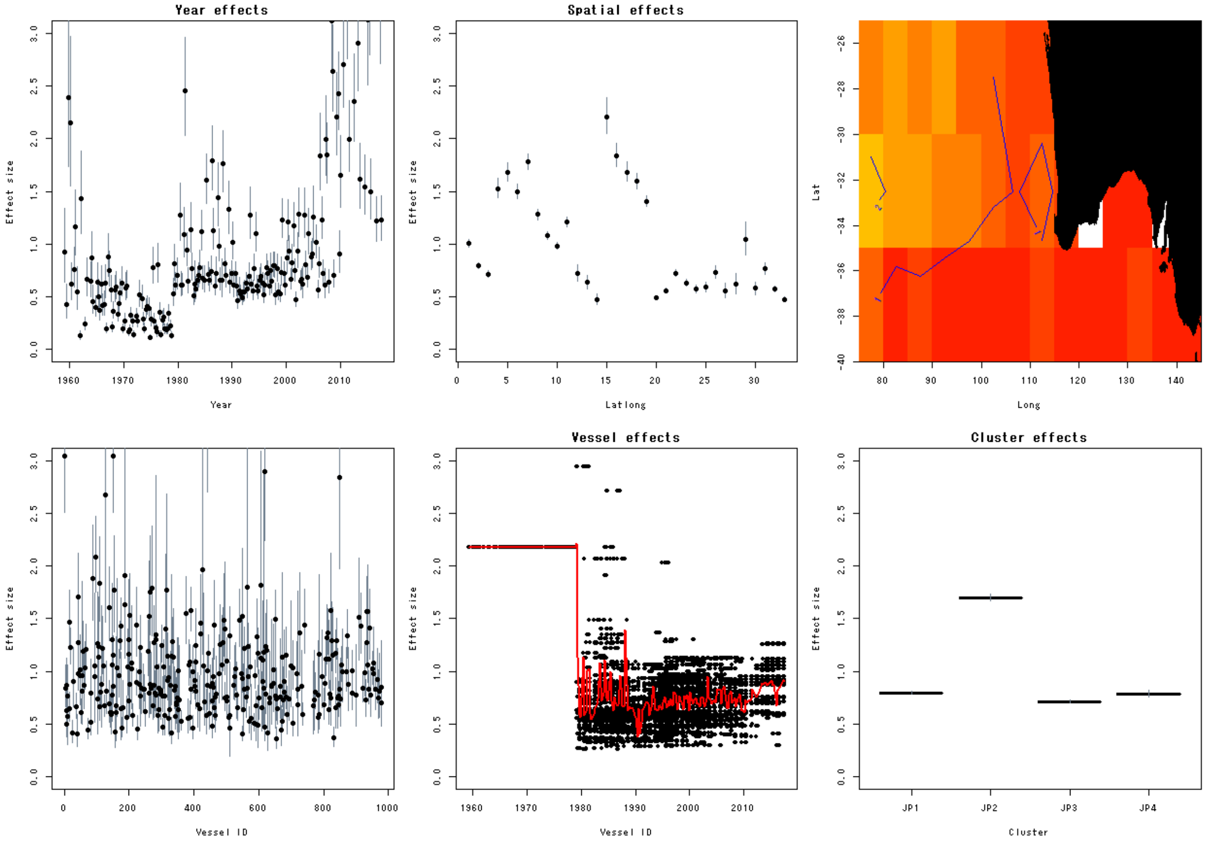


Fig. 5. Effect of each covariate for bigeye region (lognormal model). (continued)

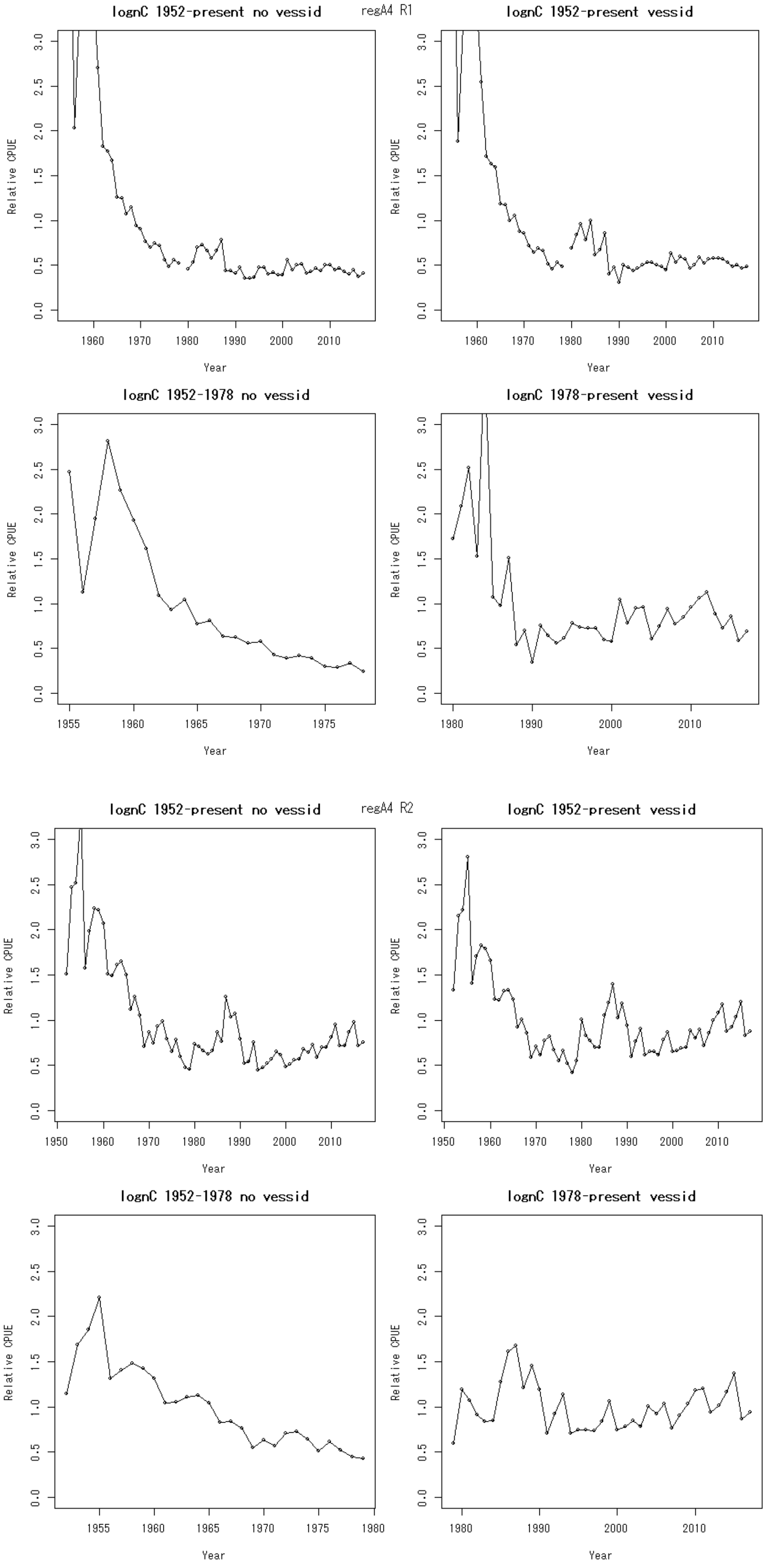


Fig. 6. Trend of annual CPUE of albacore.

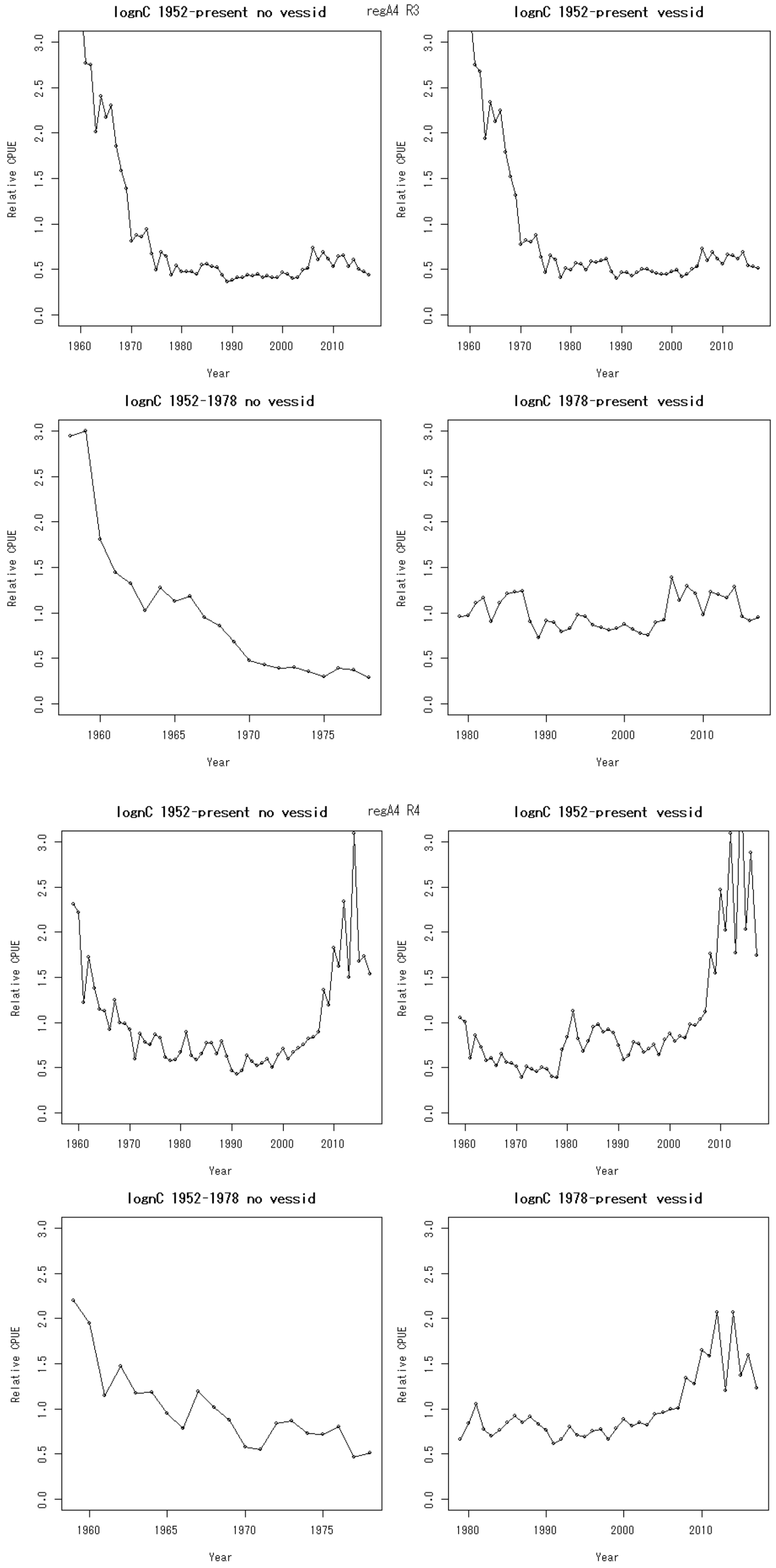


Fig. 6. Trend of annual CPUE of albacore.(continued)

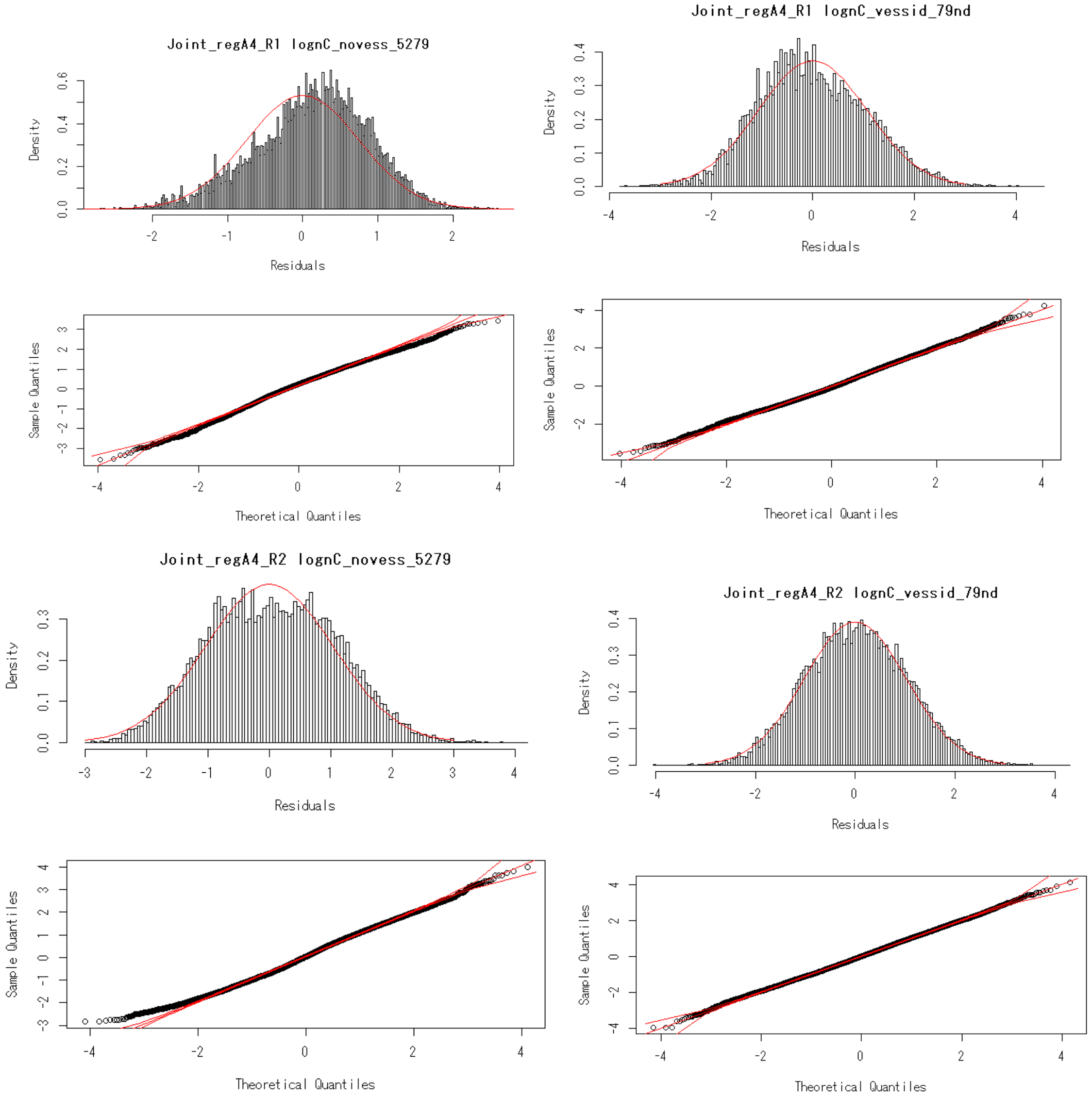


Fig. 7. Standardized residuals of CPUE standardization for albacore.

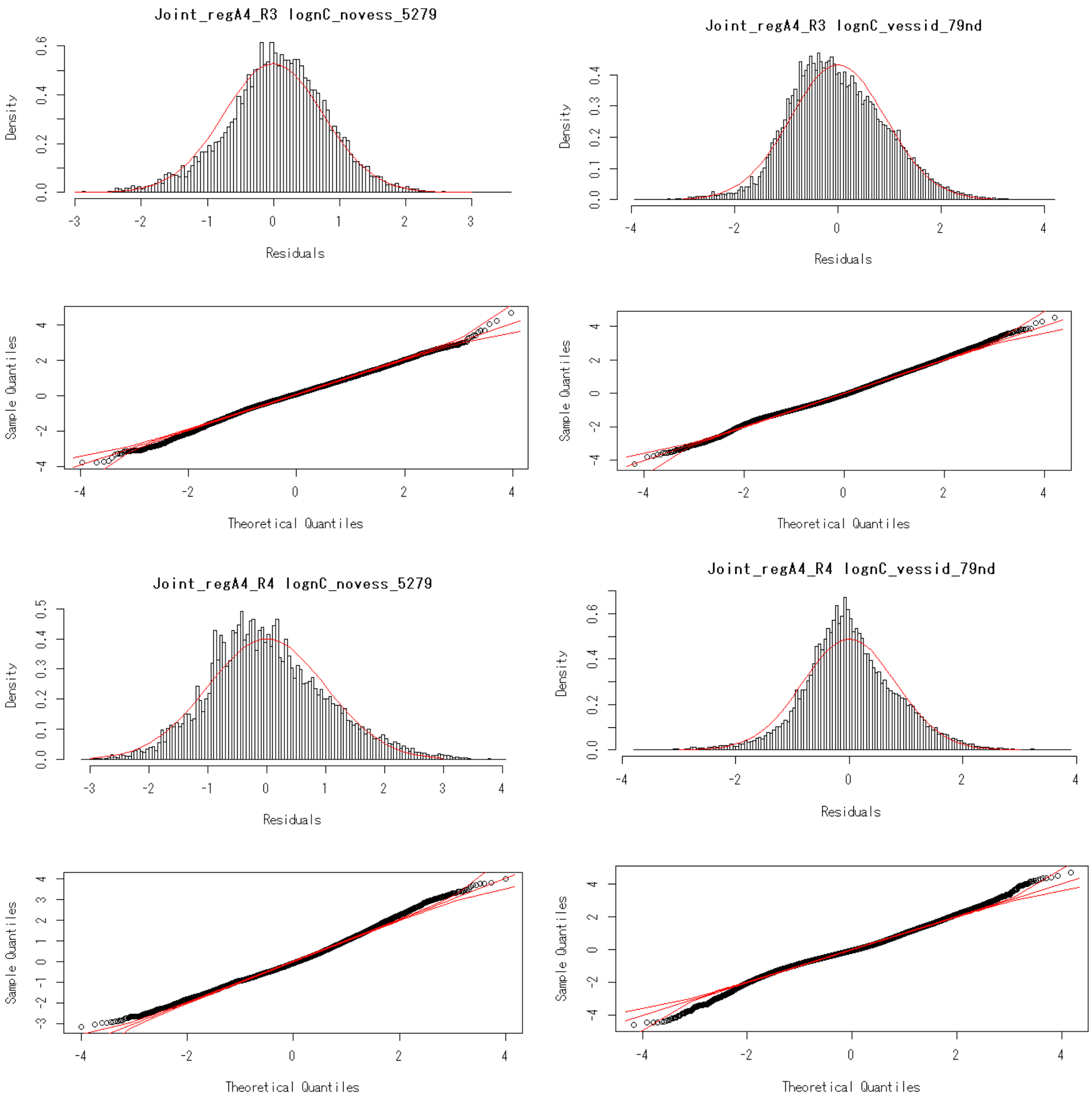


Fig. 7. Standardized residuals of CPUE standardization for albacore. (continued)