Western and Central Pacific Fisheries Commission

SCIENTIFIC COMMITTEE
TWENTY-FIRST REGULAR SESSION

Nuku'alofa, Tonga
13–21 August 2025

A comprehensive method to integrate unbiased fisheries data in spatially-explicit population dynamics models

WCPFC-SC21-2025/SA-IP-21
11 July 2025

Romain Forestier[1], Lucas Bonnin[1], Inna Senina[1], Tiffany Vidal[1], Marc Ghergariu[1], Simon Nicol[1]

[1] Oceanic Fisheries Programme of the Pacific Community

Contents

1 Executive summary

Stock assessment models and other quantitative models rely heavily on fishery-dependent data, particularly in regions where fishery-independent data (e.g., scientific surveys) are unavailable. However, the relationship between catch-per-unit-effort (CPUE) and fish abundance is impacted by variations in catchability and selectivity across different fishing operations. This paper presents a comprehensive methodology for preparing unbiased fisheries data for use in spatially-explicit population dynamics models such as SEAPODYM, a spatiotemporal model of population dynamics with age structure. Our approach addresses two key challenges: first, by systematically grouping fishing data into distinct fisheries with consistent catchability and selectivity patterns, and second, by leveraging high-resolution spatial data to maintain linear relationships between catch and biomass density at the grid cell level. We demonstrate this methodology using operational longline data from Pacific Island countries and distant-water fishing nations targeting yellowfin tuna in the Pacific Ocean. The approach incorporates covariates such as hooks between floats and target species to account for fisherman-driven changes in catchability, while assuming remaining variability is driven by environmental factors and the heterogeneity of the population density explicitly accounted in SEAPODYM. By combining these operational data with coarse resolution aggregated data from all gears, we ensure comprehensive coverage of fishing mortality while integrating fine-scale spatial resolution data needed for parameter estimation. This methodology represents a significant advancement in preparing fisheries data for spatially-explicit stock assessment models, potentially improving the accuracy of population dynamics estimates.

2 Introduction

Stock assessment models require reliable indices of fish abundance to inform model parameters and eventually assist management decisions. While fisheries-independent surveys would ideally provide these indices, such surveys are often impractical or cost-prohibitive, particularly for highly migratory species like tuna inhabiting vast regions of the Pacific Ocean. Consequently, fishery-dependent data, specifically catch and effort data, serve as crucial inputs for stock assessments.

However, raw catch and effort data present several challenges for modeling. Catch per-unit-effort (CPUE), which measures the amount of fish caught for a given amount of fishing effort, is used as a proxy for abundance. However, this relationship is complicated by two key factors: the efficiency of the fishing gear to catch fish when encountered (catchability) and the capacity of different fishing gears to target fish of specific sizes (selectivity). Understanding these factors is particularly challenging as fishing operations vary across space and time due to differences in gear types, targeting strategies and environmental conditions. Such variations must be taken into account to develop reliable abundance indices, removing biases to obtain linear relationships between catch and biomass density, as failure to do so can lead to biased estimates of stock size [Maunder and Punt, 2004, Ducharme-Barth et al., 2022], a long-recognized challenge that dates at least from Beverton and

Holt [1957].

Traditional CPUE standardization methods face significant limitations when applied to longline fisheries data, including subjective expert judgment in data selection [Braccini et al., 2011], changing fishing technologies [Hamer et al., 2024], and complex environmental influences on fish distribution [Bigelow et al., 2002]. The operational longline dataset contains relatively few variables (e.g., gear characteristics) that can effectively account for multiple sources of variation in catchability and selectivity, making it challenging to isolate abundance signals from fishing practice effects. While modern statistical spatio-temporal models (e.g., VAST, sdTMB) address some of these limitations through random effects and environmental covariates, they remain fundamentally pattern-based approaches that model correlational relationships between fish distribution and environmental conditions. These limitations are particularly acute when trying to account for the spatial structure of fish populations, as conventional models generally assume a homogeneous biomass distribution within large management regions [Punt, 2019].

SEAPODYM offers a complementary mechanistic alternative that can be used alongside traditional stock assessment approaches for fisheries management [Senina et al., 2008]. As a process-based ecosystem model, SEAPODYM simulates the biological mechanisms underlying fish distribution patterns—explicitly modeling how fish movement responds to environmental gradients in temperature, oxygen, and food availability. This mechanistic foundation creates fundamentally different data requirements: rather than removing spatial, temporal, and environmental signals as ”bias” to be standardized away, SEAPODYM uses these signals as informative biomass indices that reflect the biological processes driving fish distribution.

This approach requires structuring operational fisheries data to align with biological processes rather than statistical patterns. Fisheries must be defined not just by gear type and fishing strategy, but by how they interact with age-structured populations responding to environmental forcing. The methodology presented here demonstrates how to prepare operational longline data for SEAPODYM and can be adapted for other mechanistic ecosystem models that require process-based rather than pattern-based data preparation.

The fishing fleets targeting yellowfin in the Pacific Ocean comprise mainly two fishing gears - longlines and purse seine. There are also pole-and-line and other gears (e.g., handline, troll, ringnet) but they represent less than 15% of the catches (WCPFC Tuna Yearbook). Historical data from these fisheries were provided by the Pacific Community (SPC) for the Western and Central Pacific Ocean, and by the Inter-American Tropical Tuna Commission for the Eastern Pacific Ocean. Our approach here focuses on the use of operational fishing data from Pacific Island countries (PICs) and distant-water fishing nations (DWFNs) for longline gear. These operational data, derived from captain and observer logbooks, provides the most detailed information about fishing activities at fine spatiotemporal scales.

The operational longline data includes several covariates that impact CPUE such as the number of hooks between floats [Bigelow et al., 2002, Hoyle and Maunder, 2006] and the species composition in catches, allowing us to derive insights about the target species [Braccini et al., 2011]. These covariates are crucial as they allow us to remove fisher-driven impacts on catchability, such as changes of target species or fishing strategy, by creating fisheries defined by a single selectivity function and a catchability coefficient that is allowed to increase/decrease linearly with time [Senina et al., 2018]. After accounting for these factors, we assume that the remaining variability in catchability is driven by the spatial distribution associated with environmental variability and fish density distributions, which are explicitly described by the model.

While operational data provides detailed information at fine scales, albeit incomplete, it is essential to complement it with coarse resolution aggregated data to account for total fishing mortality to ensure that the geo-referenced dataset corresponds to the total annual stock removal.

The paper first describes the preparation of catch and effort data, followed by the treatment of length frequency data, and finally explains how these datasets are structured into fisheries for SEAPODYM.

3 Effort and catch data

We analyzed two longline fisheries datasets: (1) operational data from captains' logbooks and observer reports, and (2) a coarser resolution spatially and temporally aggregated dataset raised to total catches (hereafter referred to as ”raised data”). The study region encompasses both the Western and Central Pacific Ocean (WCPO) and Eastern Pacific Ocean (EPO). Both datasets included comprehensive fishing operation details as described in Table 1. Catch documentation differed between datasets: the operational data recorded both number of fish and weight in kilograms, while the raised data reported catches in metric tons only. We categorized the caught biomass into five groups: yellowfin, skipjack, bigeye, albacore, and ”other” (comprising all species not included in the first four categories). Rather than focusing solely on yellowfin biomass, we considered the total catch composition per vessel as it provides insights into targeted species and the catchability/selectivity patterns of each fishing trip.

3.1 Dataset formatting

The datasets have different formats, particularly in geographical coordinates, catch measurements, and effort units. We standardized all datasets using the format in Table 1. After standardization, we removed records missing geo-spatial coordinates and catch number. The spatial extent of the data was then limited to the study region (100°-290°E;61°N,59°S).

The operational datasets required particular attention for outlier removal (e.g., unreasonable values). For longline data, catch weights were often estimated or missing (as the required reporting metric is numbers), leading us to primarily rely on catch numbers as the metric of abundance. Table 1 details the data used and retained after processing, while Table 4 summarises the data removed.

Table 1: Summary of the available effort catch data and proportion of reliable data.

	Captain's data			
	Input		Processed	
	number/range	unit/format	number/range	unit/format
Total	5576014	records	99.75 %	-
Catch	100 %	kg	99.75 %	mt
Effort	100 %	count	99.75 %	hundred
Hooks Between Floats	67.9 %	count	67.8 %	count
Countries	32	flag	100 %	flag
Vessels	11164	id	99.97 %	id
Date Range	1960-06-05 - 2024-01-15	yyyy-mm-dd	1960-06 - 2022-12	yyyy-mm
Coordinates	0°-360°E;90.5°N,82.5°S	DDDMM	100°-290°E;61°N,59°S	1°x 1°
	Observer's data			
	Input		Processed	
	number/range	unit/format	number/range	unit/format
Total	272896	records	94.6 %	-
Catch	100 %	kg	94.6 %	mt
Effort	95.88 %	count	94.6 %	hundred
Hooks Between Floats	95.51 %	count	94.6 %	count
Countries	26	flag	100 %	flag
Vessels	2130	id	98.3 %	id
Date Range	1980-12-15 - 2024-02-04	yyyy-mm-dd	1987-06 - 2022-12	yyyy-mm
Coordinates	90.7°-348.5°E;83.5°S,45.65°N	DDDMM	100°-270°E;46°N, 48°S	1°x 1°
	Raised			
	Input		Processed	
	number/range	unit/format	number/range	unit/format
Total	367619	records	100 %	-
Catch	100 %	mt	100 %	mt
Effort	100 %	hundred	100 %	hundred
Countries	31	flag	100 %	flag
Fleets	29	id	100 %	id
Date Range	1950-06 - 2023-09	yyyy-mm	1950-06 - 2022-12	yyyy-mm
Coordinates	0°-360°E;70°N,85°S	5°x 5°	102°-283°E;61°N,53°S	5°x 5°

We applied statistical filtering to the effort values, in this case, number of hooks per set, removing values beyond three standard deviations from the mean of the entire dataset. Records without any reported effort could not be used in the analysis and were put aside to be used as a global catch removal variable.

The same filtering was applied to the hooks between floats (HBF) measurements, resulting in a working range of 2 to 50 hooks. In cases where records were only missing HBF data (32% of captains' logbooks and 0.3% of observer records), the records were still kept and missing values were imputed later on.

Merging Operational Datasets

The captains' logbooks and observer entries were merged using vessel flag, vessel name, and date as matching criteria. When both captain and observer data existed for the same vessel and date, observer entries were given precedence.

Spatial Processing

As previously mentioned, the data underwent spatial filtering, using a simple landmask, removing data points outside of the study region. During this process, we aggregated the operational data to 1°x 1°degree resolution, matching the same grid of cells as the raised data. Additional filtering was performed using an ERA5-derived temperature mask [Hersbach et al., 2020], which excluded all entries from SST below 10° [Weng et al., 2017].

3.2 Dividing datasets

We divided all catch-effort datasets into three categories based on their effort and yellowfin catch values. Entries with positive effort formed our primary dataset for generating fisheries files. Entries with both zero catch and zero effort were excluded from further analysis. Entries with yellowfin positive catch but zero effort were retained solely to inform fisheries mortality calculations. This categorization ensured that only relevant data points contributed to our analyses while maintaining comprehensive information for mortality assessments.

3.3 Data recovery

While operational datasets provide fine-scale spatial information, they incompletely represent longline fisheries history. Countries such as Japan, Korea, Indonesia and Taiwan provide only a portion of their operational data (due to reporting requirements or availability), creating significant gaps between the total catch of the operational data and the catch represented in the raised data. We required 1°x 1°spatial resolution data for these analyses; however, the available raised datasets were at 5°x 5°resolution. To address this discrepancy, we created a complementary ”delta longline” dataset. This dataset represents the difference between raised and operational data, degraded to 5°x 5°, calculated as:

$$V^{\Delta}_{f,g,t} = V^{R}_{f,g,t} - V^{O}_{f,g,t}$$

Where:

$V^{\Delta}_{f,g,t}$ represents the delta longline dataset

f indicates the flag

g represents the 5°x 5°grid cell (latitude, longitude)

t is the time (month)

$V = \{C, E\}$ where C represents catch and E represents effort

$V^{R}_{f,g,t}$ is the raised data

$V^{O}_{f,g,t}$ is the operational data (captain's and observer's)

We first aggregated the operational data to match the raised data's temporal (monthly) and spatial (5°x 5°) resolution, grouping by nation. The resulting delta longlines dataset was then used in conjunction with the operational data to provide a complete historical representation of longline fishing activities. In some cases, the operational data catch were higher than the raised data, resulting in negative values, which were changed as 0.

4 Yellowfin length frequency data

We analyzed two length frequency datasets: (1) operational data from observer and port sampling trips, and (2) a regionally aggregated dataset (hereafter referred to as "aggregated LF data"). The operational data contains two distinct collection methods: observer sampling, which records fish lengths with precise coordinates and catch times during fishing days, and port sampling, which measures representative catch samples at trip conclusion, documenting only the start and end coordinates of the trip alongside sampling time. The aggregated LF data incorporates the operational data alongside historical data and Japanese submissions to the Western and Central Pacific Fisheries Commission (WCPFC). These aggregated data are often consolidated into large regions ranging from 5°x 5°to 20°x 10°grid cells by quarter. While lacking the spatial precision of operational data, the aggregated dataset fills the temporal and spatial gaps in the operational records. Comprehensive details of both datasets are described in Table 2.

4.1 Dataset formatting

Similar to the catch-effort data, the first step was to standardise the format for geolocation and time described in Table 2, followed by the removal of entries containing null values for key data fields or being outside of the study region. An exception was made for the length and length code of the operational data where we attempted to salvage as much data as possible, even when considered

Table 2: Summary of the available length frequency data before and after processing. A sample encompass all the length frequency measurement during a day from one vessel for the detailed data and in one region for the aggregated data.

	Port sampling			
	Input		Processed	
	number/range	format/bin	number/range	format/bin
Samples	15982	1 cm	82.2 %	1 cm
Countries	17	flag	94.1 %	flag
Vessels	2013	id	20.6 %	id
Date Range	1994-04-05 - 2024-02-08	yyyy-mm-dd	2002-01-11 - 2022-12-31	yyyy-mm-dd
Coordinates	0°-257.5°E;25.2°N,27°S	DDDMM	121°-224°E;24°N,29°S	1°x 1°
	Observer			
	Input		Processed	
	number/range	format/bin	number/range	format/bin
Samples	99104	1 cm	92.5 %	1 cm
Countries	24	flag	100 %	flag
Vessels	1272	id	99.53 %	id
Date Range	1980-12-15 - 2024-02-04	yyyy-mm-dd	1980-12-15 - 2022-12-31	yyyy-mm-dd
Coordinates	95.9°-267.1°E;44.7°N,49.7°S	DDDMM	100°-267°E;44°N,50°S	1°x 1°
	Aggregated			
	Input		Processed	
	number/range	format/bin	number/range	format/bin
Samples	33141	1,2 and 5 cm	99.7 %	1 cm
Countries	29	flag	100 %	flag
Fleets	6	id	99.97 %	id
Date Range	1948 - 2023	quarter	1948 - 2022	quarter
Coordinates	0°-360°E;40°N,85°S	5°x 5°; 10°x 5°; 20°x 10°	100°-280°E;40°N,50°S	5°x 5°; 10°x 5°; 20°x 10°

NULL, which is detailed below.

The operational size data encompass all gears but the majority of the data comes from:

- port sampling: length data are obtained by randomly sampling the trip at offload. These data are only recorded 10-20% of the time, only provided by the PICTs and apply primarily to the longlines fishery. There is usually no coverage for vessels fishing outside of PICTs waters. Each port sampling is assigned a unique trip ID.
- longline observers: length data are obtained by measuring the length of every fish being brought onboard , but they are not present on every trip (5% coverage of all vessels fishing in the study area). All the entries from a single fishing trip are assigned a unique trip ID.

4.2 Data recovery

The length reported in the operational data are associated with a reported length measurement type code. The ”UF” (upper jaw to caudal fork) length code was picked as the reference as it is commonly used. Using Macdonald et al. [2022], we found that other length codes could be converted to ”UF”.

$$UF = 3.951SD^{0.8369}$$

$$UF = 11.385PS^{0.6619}$$

$$US = SD$$

Some length codes were recorded as NULL. In these cases, the unique trip id was used to determine if there were other samples from the same trip and species reported with a length measurement type code. If so, that code was assumed to be the correct one and was imputed in the dataset. Note that 20% of port sampling records were removed due to no length being recorded or missing coordinates.

4.3 Matching effort-catch data and length frequency data

We merged the EC and LF operational datasets to obtain a mean length associated to effort catch data, to be used as a clustering variable. The observers and port sampling data were treated separately as they captured different temporal scales - observer data provided length frequencies for daily catches, while port sampling measured length frequencies for an entire trip.

To begin, the LF observer entries are matched to the EC data using fishing event dates (in days), vessel and flag name. Data types (i.e., EC or LF) that had not matched with their counterparts were still retained. To approximate the trip structure (with the trip ID) of the LF data, fishing trips were created in the EC data by assuming that chronologically consecutive entries by a given vessel were part of the same fishing trip, allowing a maximum gap of three consecutive days without entries within the same trip.

Using these defined trips, we matched EC data with port sampling entries that fell within the temporal bounds of each EC trip. Since port sampling occurs only on the last day of a fishing trip, when a port sampling record was identified within an EC trip, all EC data entries up to and including the port sampling date were assigned to that trip. When multiple port sampling entries occurred within what was initially defined as a single EC trip (based on the three-day gap rule), this indicated that the original EC trip definition was incorrect. In such cases, the EC trip was subdivided into separate trips, with each new trip ending on a port sampling date, regardless of the three-day gap rule. An average fish length for each trip was calculated based on the port sampling entries and attributed to each day of the trip without replacing any existing observer data, if present. Furthermore, fish counts from port sampling were allocated to individual days proportionally based on the daily catch entries in the EC data. This allocation approach assumes that the length frequency distribution from port sampling is representative of the entire trip's catch, with the proportion of measured fish assigned to each day reflecting the relative fishing catch rate documented in the EC data.

Following this process, 2.4% of EC data daily entries were assigned a mean length from observer or port sampling measurements. When a direct measurement was not available for a daily EC data entry, but data were available for other days of the same trip, an average of the available length data was used as a proxy to fill the gap. The EC data was then aggregated monthly at 1°x 1°per vessel and flag. The average length for each monthly entry was calculated from the original LF data if available, otherwise using the proxy length.

After this process, 4.5% of the aggregated EC data entries were associated with a mean length. To complete the missing entries, we used the aggregated LF data. This dataset was organized in longitude-latitude grid regions ranging from 5°x 5°to 20°x 10°and was aggregated by quarters, years, and flags. We matched the EC entries, the ones occurring in the same temporal period (year/quarter) and with identical flags, within the LF regions, specifically targeting EC entries that had not already received length frequency assignments from the operational data. This matching meant that for a given date and flag, multiple 1°x 1°grid cells of EC data falling within a larger LF region would be assigned that region's mean length. In instances where multiple LF regions overlapped, the assigned mean length was calculated as the weighted mean of the LF regions, using their respective fish counts as weights.

The final dataset was aggregated to a monthly resolution, containing values for catch, effort, and mean length (when available) per vessel and flags.

The delta dataset of longline fisheries was merged with the aggregated LF data the same way as with the operational data, which yielded 30% coverage of the EC data.

5 Defining fisheries

5.1 Clusters method

The following clustering method was aimed at partitioning fishing operations into distinct subsets, characterised by similar fishing technique and selectivity (Figure 1). The operational and delta datasets were partitioned separately to obtain clusters that will then be used as fisheries with single catchability and selectivity parameters within the SEAPODYM modelling framework. This aimed to create homogeneous groups of fishing operations that reflect similar fishing practices and effectiveness, to enable the use of linear relationship between tuna biomass density and catch within each geo-location and time.

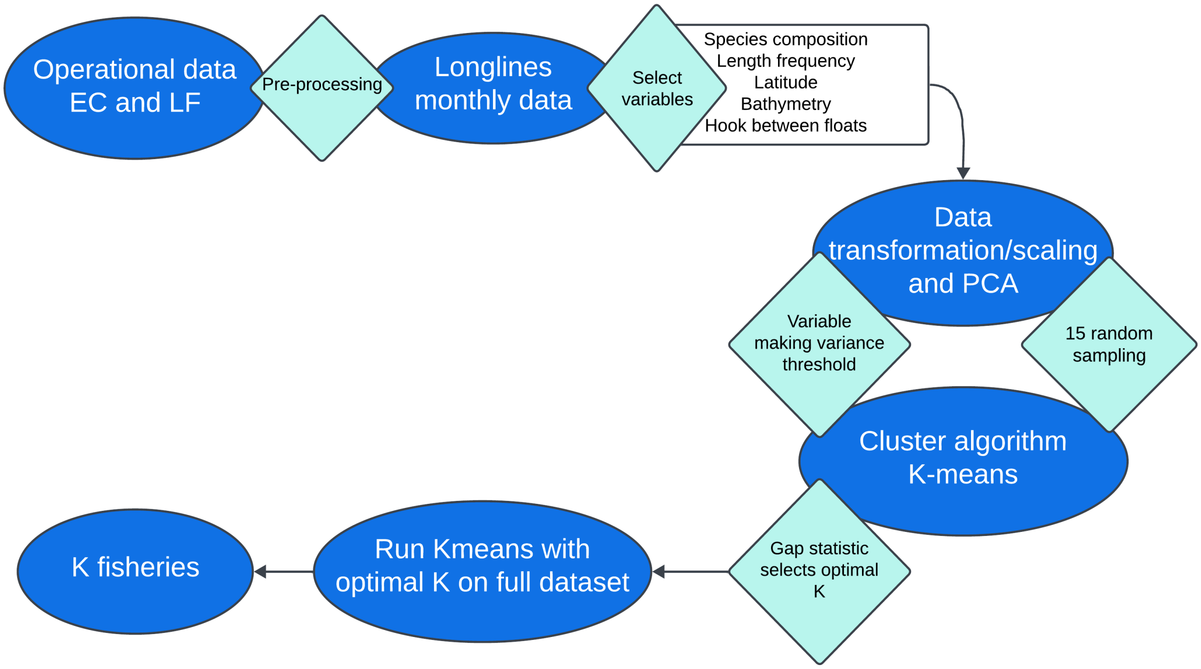

Figure 1: Flowchart of the clustering methodology used to partition longline fishing operations into homogeneous fisheries.

5.1.1 Operational data

Clustering was performed using available covariates potentially indicative of fishing strategies such as species composition, number of hooks between floats, and the mean length of the catch . Additional stratification factors included latitude (rounded to 1 degree) and water biomes. Three water biomes were defined based on the depth of the water column and steepness of the ocean floor in each grid cells: the continental shelf, characterized by gentle slopes (less than 3 degrees) and shallow waters (above -200 meters); the continental slope, distinguished by its steeper gradient (more than 3 degrees) but only above - 1000 meters; and the abyssal plains, which feature gentle slopes (less than 3 degrees) in the deep ocean (below -200 meters) or any slopes past - 1000 meters.

To maximize the use of catch-effort (EC) data, entries with missing mean length measurements (52%) and hooks between floats (HBF; 32%) were imputed through Multiple Imputation by Chained Equations (MICE) [Van Buuren and Groothuis-Oudshoorn, 2011]. Due to the low coverage of length frequency data, we assume that at a monthly level, the missing values can be determined using the existing data coupled to the species composition of the catch as predictor variables. To facilitate the clustering analysis, HBF was discretized. We employed model-based clustering (Mclust package in R) to identify natural groupings within the distribution of HBF, the optimal model of four groups was selected using the Bayesian Information Criterion. Clustering was performed using data from 1990 to 2022 only, which was the most complete data, considering yellowfin, bigeye, albacore and other in species composition of catch.

We tested different combinations of the covariates at hand (i.e., scenarios) and calculated several metrics (described below) to be able to rank them and select the most sensible fisheries. Given the heterogeneous nature of the dataset, variables were standardized using Z-score normalization to ensure comparable scales and then a principal component analysis (PCA) was applied to effectively reduce high-dimensional data to a lower-dimensional space while preserving the maximum amount of information in the data [Hotelling, 1933]. Principal components cumulatively explaining at least 70% of the total variance were retained for clustering analysis. This threshold-based selection of components effectively eliminates dimensions contributing minimal variance, thereby reducing noise while preserving the dataset's primary structure. This dimensionality reduction enhances computational efficiency of the clustering algorithm. In order to cope with the large size of the dataset (e.g., 2e6 entries for the operational data), clustering was performed over random samples from the dataset. Fifteen random samples the size of 1% of the total data were used for the clustering. 1% was enough to describe the structure of the entire dataset by assessing the consistency of the final clusters resulting from the 15 iterations. The subsets were subjected to K-means clustering analysis [Xie and Jiang, 2010] to determine the optimal number of clusters, with iterations testing cluster numbers ranging from 2 to 20 (empirically, iterations resulting in more than 20 clusters were never observed). Optimal cluster number K was determined using the Gap statistic method [Tibshirani et al., 2001], which evaluates the change in within-cluster dispersion against that expected under a null reference distribution. The smallest value of K such that the gap statistic is within one standard deviation of the gap at K+1 was used.

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_k + 1$$

Gap statistics performed over the 15 subsets were ensured to return the same optimal number of clusters (if not, new subset were drawn until 15 agreed on the optimal number of clusters). Each of the subsets were matched by their similarities and their centroid averaged to obtain a mean centroid value and its normalised root mean square error for each cluster, which was then used to assess the stability of the results. This approach distinguished between unstable clustering solutions (where centroids varied across random samples) and stable solutions (where centroids remained consistent across random samples).

K-means was then run on the entire dataset with the selected K and a final check was made to ensure that each cluster contained at least 3% of yellowfin total catch. If not, clusters were removed and the data were assigned to the closest centroids left. Furthermore, the silhouette concept [Rousseeuw, 1987] was used to assess the quality of the clustering. Silhouette width is a metric indicative of how similar an entry is to its own cluster, compared to other clusters. It ranges from -1 to 1 and scores higher when clusters are dense and well separated. This metric further helped flag scenarios with bad cluster quality.

Finally, a breakpoint analysis was conducted on the time series of variables (e.g., CPUE as catch unit per hundred of hooks, length) to assess the consistency of the data clustered through time. This method detects ecologically meaningful regime changes in fisheries time series through a multi-step approach designed to identify transitions between stable operational states rather than temporary fluctuations. First, the method filters data to the post-1990 period to focus on the modern fishing era and applies loess smoothing (span = 0.3) to reduce short-term noise while preserving genuine regime transitions. The core detection employs the PELT (Pruned Exact Linear Time) algorithm [Killick et al., 2012] with BIC penalty to identify points where mean levels change significantly, using a minimum segment length of 12 months to prevent over-segmentation. Each detected segment is then classified as ”stable,” ”increasing,” or ”decreasing” through linear regression analysis, where segments are labeled as trending only if they meet both statistical significance (p-value ¡ 0.1) and practical significance (slope magnitude ¿ 10% of within-segment standard deviation) criteria, ensuring that detected trends represent genuine patterns rather than random variations. Finally, the method validates regime changes by examining transitions between stable periods, retaining only breakpoints where the change in mean level exceeds a specified threshold (typically 3%) and passed the 2 criteria described below, effectively filtering for meaningful shifts while ignoring minor fluctuations.

- Segment Length Criterion: For biomass dependent variables such as CPUE and mean length in catch, segments shorter than the mean of the yellowfin age distribution currently fished in the time series (3.5 years on average) were initially flagged as potential breakpoints. Abrupt change in CPUE or length that occurs in a time shorter than the population renewal of yellowfin were not considered as coming from a change in the yellowfin's biomass but from how the data were clustered instead.

- Spatial Dispersion Criterion: For each potential breakpoint, we evaluated the fishing fleet centroid movement, dispersion around the centroid and overlap of fishing area at the start and end of the segment. A change in these metrics within the transition period between two breakpoints indicated fleet relocation, which would explain the change in the variable and negate the breakpoint classification.

This methodology allowed for robust identification of meaningful temporal shifts in the fisheries data, distinguishing between genuine structural changes and mere spatial rearrangements of the fishing fleet.

These three metrics were employed in identifying the best scenarios for K-means clustering. The goal was to determine the most stable configuration, with an ideal low nRMSE and high silhouette for each cluster per scenario and zero breakpoints indicating consistent characteristics across the clustered data. This helped narrow down the possibilities until a final manual check to assess the realism of the clusters.

5.1.2 Delta data

The method to cluster the delta data was the same, albeit with variables limited to the species composition between yellowfin, bigeye and albacore, alongside latitude and mean length.

5.2 Outlier Hampel filter

CPUE outlier values were filtered using a Hampel filter with a spatio-temporal rolling window. For each cluster, each data point was evaluated within the context of surrounding observations falling within a 250-kilometer radius and a temporal window that included the same calendar month across a ±5 year period centered on the observation date. The median absolute deviation (MAD) was calculated from this local subset of data to establish robust measures of central tendency and dispersion. This adaptive approach accounts for spatial variations in catches while ignoring seasonal variations. When outliers were identified, rather than removing the data points, effort values were adjusted to the maximum threshold permitted by the Hampel filter (defined as the median ± 7 × MAD). This conservative approach preserves catch information while moderating the influence of extreme effort values that could distort CPUE calculations.

Length frequency outliers were filtered out using a similar spatio-temporal Hampel filter based on the mean length values. The study area was partitioned into 20°x 10°grid cells and within each grid cell, the length distribution was analyzed independently, recognizing that fish size structures can vary substantially across different oceanographic regions. The Hampel filter was applied using the MAD criterion, where length observations exceeding seven times the MAD from the regional median were classified as outliers and subsequently removed from the dataset ($< 0.1\%$ removed). This regionalised approach ensures that outlier detection is sensitive to natural geographic variations in fish size distributions while identifying biologically improbable length measurements that may result from measurement or recording errors.

6 Cluster results

6.1 Evaluation of Clustering Quality

We systematically examined multiple variable combinations and dimensionality reduction thresholds to identify optimal clustering configurations. The principal component analysis (PCA) variance threshold had a strong impact on the resulting number of clusters as well as the quality of these clusters. A quantitative assessment of cluster quality was made using all variable combinations, cluster stability (which is measured using nRMSE of cluster centroid coordinates), breakpoint analyses , and silhouette values. For operational data, the lowest nRMSE scores, and therefore highest centroid stability occurred on average for the lowest variance threshold of 50% with the stability decreasing with increasing variance threshold (Figure 4). However, this trend was not necessarily true for all scenarios, with some scoring low nRMSE for any variance threshold (e.g., yba.hbfC.len.lat in Figure 10). The number of breakpoints was not significantly affected by the variance threshold (Figure 2) with an average of 2 breakpoints per cluster for HBF, 1 for mean length and 0 for CPUE. Finally, the highest silhouette scores occurred for lower variance threshold (50 and 70 %)(Figure 3) but when examining individual scenarios, high silhouette score could still be achieved with a 100% variance threshold (e.g., sp in Figure 12). We selected valid scenarios first based on their nRMSE score, then their number of breakpoints per variable (with priority on CPUE, then length and finally HBF) and finally with a reasonably high silhouette value. Scenarios producing only one cluster or no clusters with a majority of yellowfin were discarded. A preference was also given to scenario considering only the species yellowfin, bigeye and albacore in case of a tie with scenarios also including the ”other” species category, to be able to use the same variables as with the delta data which only includes the 3 species mentioned. Finally, this reduced the number of adequate scenarios to a handful that could be visually inspected for breakpoints missed by the method and spatial repartition matching the known whereabouts of the species composition within each cluster. Following these criteria, we selected the scenario based on the proportion of targeted species in the catch (yellowfin, bigeye, and albacore) and latitude using the 70% variance threshold (same result as with 50 % variance). This configuration yielded three distinct clusters, each with silhouette values exceeding 0.4, with one cluster surpassing 0.6—indicating good separation and cohesion. The corresponding nRMSE values remained consistently low (¡0.05), demonstrating high centroid stability across all 15 iterations and the breakpoint analysis showed no breaks.

For the delta dataset, the optimal clustering was achieved using only the three target species proportions without latitude. This configuration maintained high silhouette values and low nRMSE while producing no CPUE breakpoints, indicating stable fishery definitions through time.

We specifically examined the impact of including the ”other” species category in our clustering variables. In most scenarios, its inclusion generated an additional fourth cluster while simultaneously reducing silhouette values across all clusters, suggesting decreased cluster cohesion. When restricting the analysis to only the four target species categories (yellowfin, bigeye, albacore, and ”other”), the algorithm formed three clusters, but problematically combined yellowfin and bigeye into a single cluster. This configuration was deemed unsuitable for our objectives since distinguishing between these commercially important species with distinct ecological niches is essential for accurate fishery characterization.

Additional variables, such as bathymetry and hooks between floats (HBF), did not consistently improve clustering metrics, suggesting that fishing strategy (as indicated by target species composition) and spatial distribution (latitude) were the most significant determinants of fishing strategy.

6.2 Characteristics of Operational Data Clusters

The operational data were partitioned into three distinct fisheries (Table 3), each characterized by different species compositions, spatial distributions, and CPUE patterns (Figure 5). While species targeting patterns were the primary determinants in cluster formation, the role of latitude, though secondary, proved crucial in optimizing cluster quality. The inclusion of latitude increased silhouette values for two clusters and eliminated length breakpoints that were present when using only targeted species variables. This importance is further supported by our mean decrease accuracy analysis (Figure 7), which reveals latitude substantially influenced the formation of clusters 2 and 3, even while remaining secondary to species composition variables.

Fishery 1: Yellowfin-Targeted Tropical Fishery

This cluster represents the primary yellowfin tuna fishery, accounting for 61% of total yellowfin catch in the operational dataset. Catch composition within this fishery is predominantly yellowfin (73%), with bigeye tuna comprising most of the remainder. Operations are concentrated in tropical waters of the Western and Central Pacific Ocean (WCPO), as evidenced by the distinct spatial pattern in Figure 5A.

The mean length of yellowfin caught in this fishery remained stable at approximately 118 cm throughout the study period, suggesting consistent selectivity targeting mature individuals. Temporally, CPUE exhibited a gradual decline until 2000, after which it stabilized through 2022, potentially reflecting changes in fishing practices or management measures implemented during this period.

Fishery 2: Bigeye-Dominated Northern Fishery

The second cluster accounted for 15% of total yellowfin catch but was primarily bigeye-targeted, with yellowfin comprising 24% of catch composition. This fishery operated predominantly in the northern hemisphere with a broader spatial distribution than Fishery 1 (Figure 5B). Yellowfin caught in this fishery showed similar mean lengths (119 cm) compared to the first cluster. CPUE followed similar temporal trends to Fishery 1 but with values 3-5 times lower, consistent with yellowfin being a secondary rather than primary target.

Fishery 3: Albacore-Dominated Widespread Fishery

The third cluster contained 24% of total yellowfin catch, despite yellowfin comprising only 11% of its catch composition, which was predominantly albacore. This fishery exhibited the widest spatial distribution, spanning the entire Pacific with concentration in the southern hemisphere (Figure 5C). Mean yellowfin length in this fishery was notably smaller (113 cm) than in other clusters, suggesting different size selectivity possibly related to shallower setting practices associated with albacore targeting. CPUE values were the lowest among all clusters and displayed more pronounced seasonality.

CPUE Distribution Patterns

Examination of CPUE frequency distributions (Figure 6) revealed distinct patterns across the three fisheries. Fishery 1, which specifically targeted yellowfin, displayed a broad, relatively uniform distribution of CPUE values, indicating variable success rates but consistent targeting behavior. In contrast, Fisheries 2 and 3, where yellowfin was not the primary target, showed highly left-skewed distributions with most observations clustered near zero, characteristic of incidental or opportunistic yellowfin capture within fisheries primarily targeting other species. Spatially, despite differences in overall CPUE magnitude and distribution, all three fisheries exhibited peak CPUE values in tropical waters (Figure 5), consistent with the known ecological preferences of yellowfin tuna. However, the spatial extent and CPUE gradients differed substantially between clusters, reflecting different targeting strategies and operational characteristics.

6.3 Characteristics of the Delta Data Clusters

The delta dataset, representing the complementary raised data, was similarly partitioned into three distinct fisheries based on species composition, with each cluster exhibiting unique CPUE patterns and spatial distributions.

Delta Fishery 4: Yellowfin-Dominated Fishery
This cluster accounted for 73% of total yellowfin catch in the delta dataset, with catch composition predominantly yellowfin (76.5%) and bigeye comprising most of the remainder. This parallels the species composition observed in Fishery 1 of the operational dataset, suggesting consistency in yellowfin targeting strategies across different data sources.

CPUE in this cluster exhibited high variability before 1980, followed by a gradual decline until 2010, after which it stabilized. Values ranged from 0.4 to 1.4, representing the highest CPUE among delta clusters. Spatial distribution revealed high CPUE concentrations in tropical waters of both the WCPO and along the coast of the EPO, with a notable area of lower CPUE in the offshore regions of the EPO, creating a distinctive gap pattern in the spatial distribution.

Delta Fishery 5: Bigeye-Dominated Fishery
The second cluster contained 15.5% of total yellowfin catch but was primarily composed of bigeye tuna. This cluster's CPUE exhibited high variability until 1980, followed by relative stability until 2000, then decreased between 2000 and 2010 before stabilizing again. CPUE values ranged from 0.4 to 0.28, substantially lower than in Delta Fishery 4.

Spatially, Delta Fishery 5 maintained the general pattern of higher CPUE values in the tropics, with a lesser contrast between coastal and offshore regions in the EPO. This suggests a more uniform distribution of fishing success across the EPO for this bigeye-dominated fishery.

Delta Fishery 6: Albacore-Dominated Fishery
The third cluster comprised 11.5% of total yellowfin catch and was predominantly composed of albacore tuna. CPUE time series exhibited strong seasonality throughout the study period, with particularly high variation before 1970. After 1970, CPUE maintained a relatively constant trend with continued marked seasonality, though with reduced amplitude. CPUE values ranged from 0.02 to 0.29.

The spatial distribution of CPUE in Delta Fishery 6 showed a distinctive pattern in the EPO, with high CPUE values concentrated along the entire western coastline of the continental United States with lower CPUE further west compared to the two other fisheries.

CPUE Distribution Patterns

The CPUE frequency distributions across the three delta fisheries (Figure 9) closely paralleled those observed in the operational dataset. Delta Fishery 4, targeting yellowfin as the primary species, exhibited a broad distribution of CPUE values similar to Fishery 1, while Delta Fisheries 5 and 6, where yellowfin represented incidental catch, displayed the characteristic left-skewed distributions with most observations concentrated near zero CPUE values. Spatially, all three delta fisheries maintained the fundamental pattern of peak CPUE values in tropical waters (Figure 8), consistent with yellowfin ecological preferences and mirroring the spatial distributions observed in the operational clusters. The primary distinction appeared in the Eastern Pacific Ocean coverage, where the delta dataset provided more comprehensive spatial representation than the operational data, particularly for the albacore-dominated fishery along the continental coastlines.

6.4 Comparison with Operational Data Clusters

The delta and operational datasets yielded remarkably similar clustering structures, with three distinct fisheries primarily differentiated by species composition. Both approaches identified discrete yellowfin-dominated, bigeye-dominated, and albacore-dominated fisheries with comparable proportions of total yellowfin catch. This consistency across different data sources reinforces the robustness of our clustering approach.

The most notable difference between operational and delta clusters appeared in their spatial CPUE distributions, particularly in the EPO, where the operational data is lacking. These differences likely reflect the complementary nature of the two datasets, with delta data capturing fishing activities not represented in the operational data entries.

6.5 Validation of Clustering Approach

The effectiveness of our clustering approach is evidenced by several key outcomes. First, the absence of breakpoints in both CPUE and length time series within the optimal clustering configuration demonstrates temporal stability in catchability and selectivity—a critical requirement for reliable abundance indices. Second, the distinct CPUE distributions and spatial patterns observed across clusters confirm that our approach successfully identified genuinely different fishery types with consistent operational characteristics.

The Hampel filter identified and adjusted 0.8% of outliers in the CPUE data. These adjustments improved within-cluster cohesion without substantially altering overall CPUE trends, suggesting that our approach successfully mitigated the influence of extreme values while preserving the underlying signal.

Table 3: Summary metrics of the fisheries

Fishery	min CPUE	max CPUE	min length	mean length	max length	min lat	max lat	% total	% YFT
1	0.40	1.38	108.09	117.67	127.13	-34.5	19.5	61.33	73.02
2	0.02	0.24	106.01	118.88	131.50	-13.5	35.5	14.73	24.30
3	0.00	0.28	103.20	113.33	124.32	-44.5	31.5	23.93	10.64
4	0.39	1.41	101.80	117.07	132.45	-27.5	32.5	72.96	76.53
5	0.04	0.28	106.62	122.47	137.01	-27.5	37.5	15.62	20.50
6	0.02	0.29	97.31	114.98	132.58	-42.5	37.5	11.42	8.06

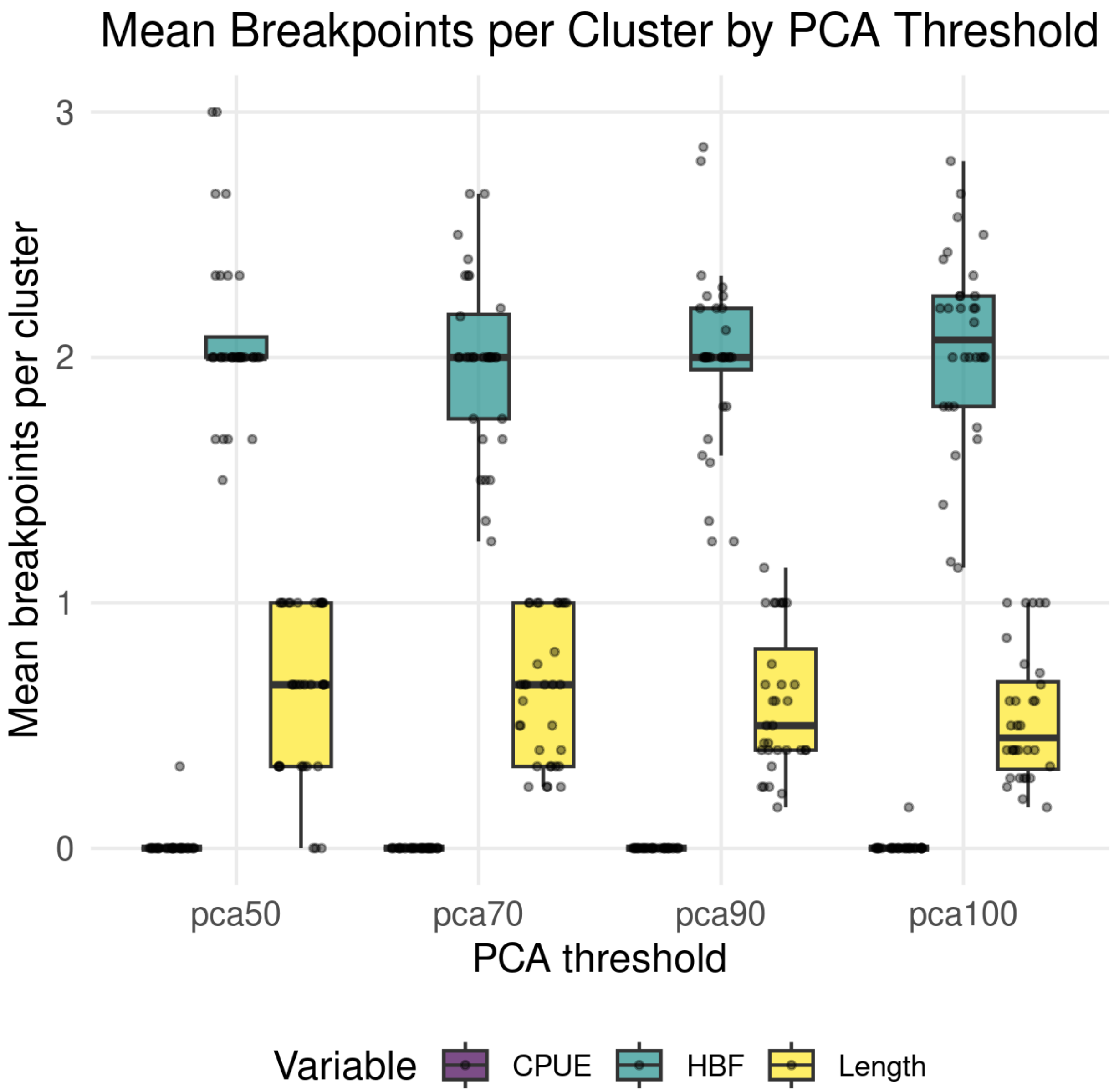

Figure 2: Boxplot showing the number of breakpoints per CPUE, length, and HBF per PCA method. Each method differs by the amount of variance kept to do the K-means clustering, either 50, 70, 90 or 100% (x-axis). Each bar is the average amount of breakpoints across all tested scenario weighted by their number of clusters.

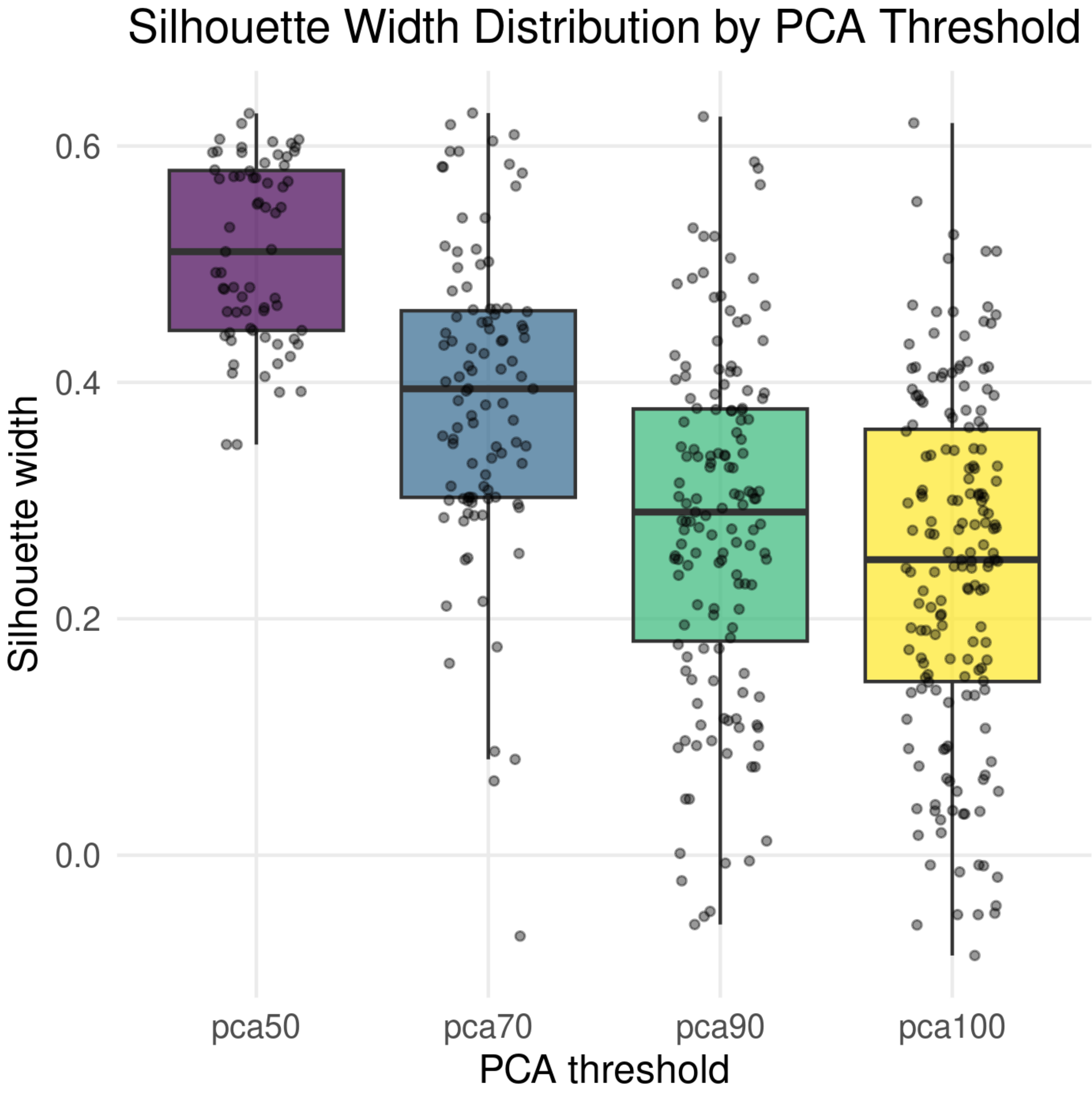

Figure 3: Boxplot showing the averaged silhouette width per cluster for each PCA method.

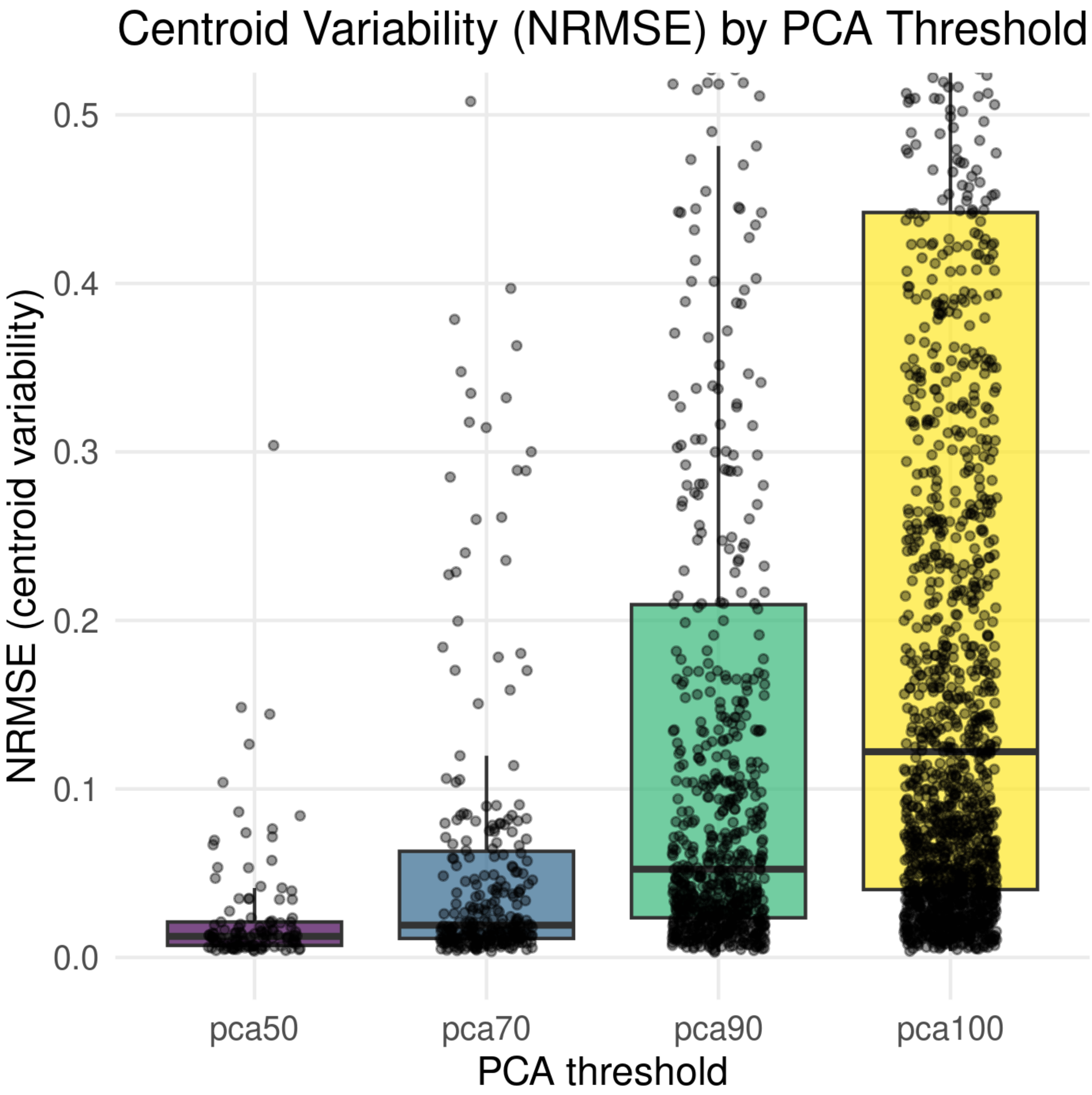

Figure 4: Boxplot showing the normalised root mean square error for each PCA method. Some values were to high too be showed entirely while conserving a scale allowing to see the small values and were cut at 0.5.

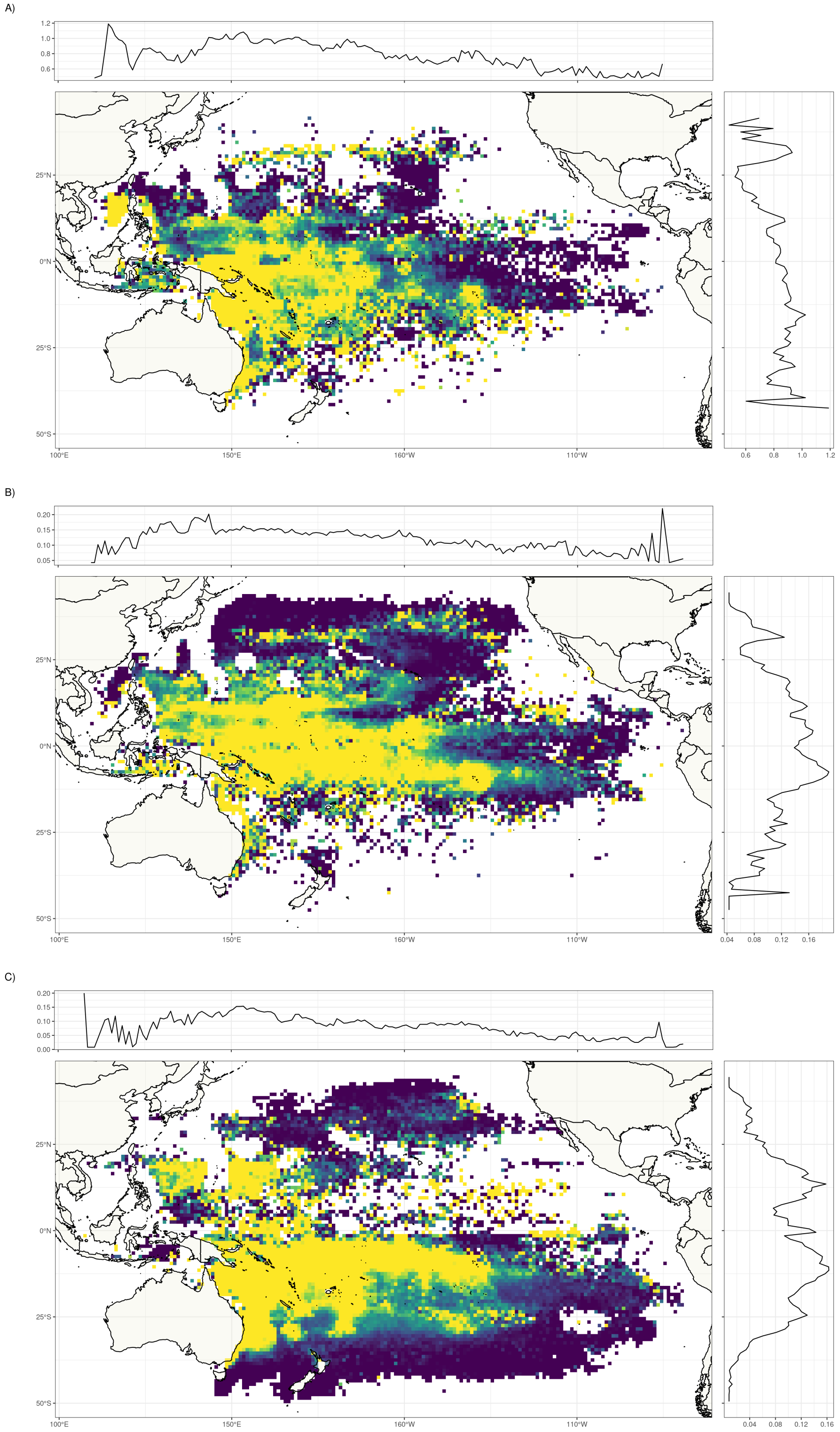

Figure 5: Maps showing the spatial range of the fishery and its average CPUE over time (1960-2022). The curves at the top and right of the maps shows the CPUE average value per longitude and latitude respectively. A through C describe each fishery from 1 to 3 respectively.

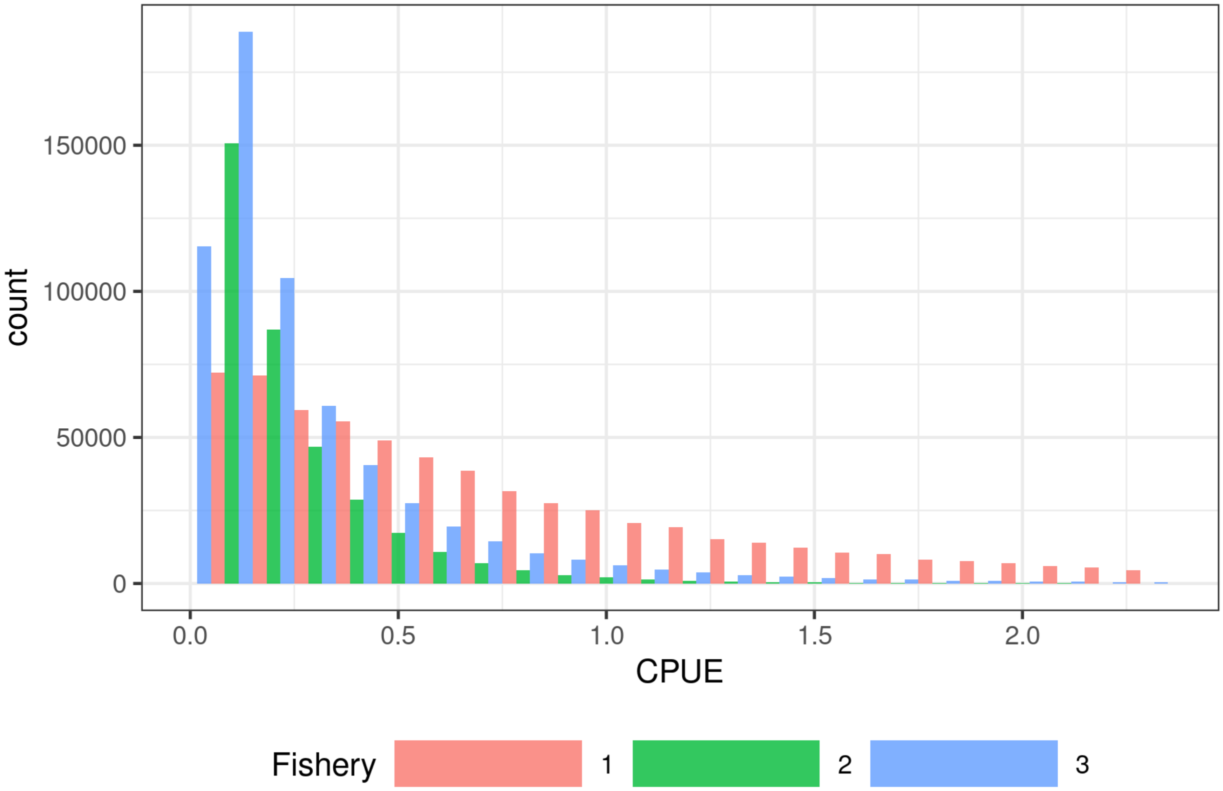

Figure 6: Frequency distribution of CPUE values (yellowfin catch per hundred hooks) across the three operational fisheries.

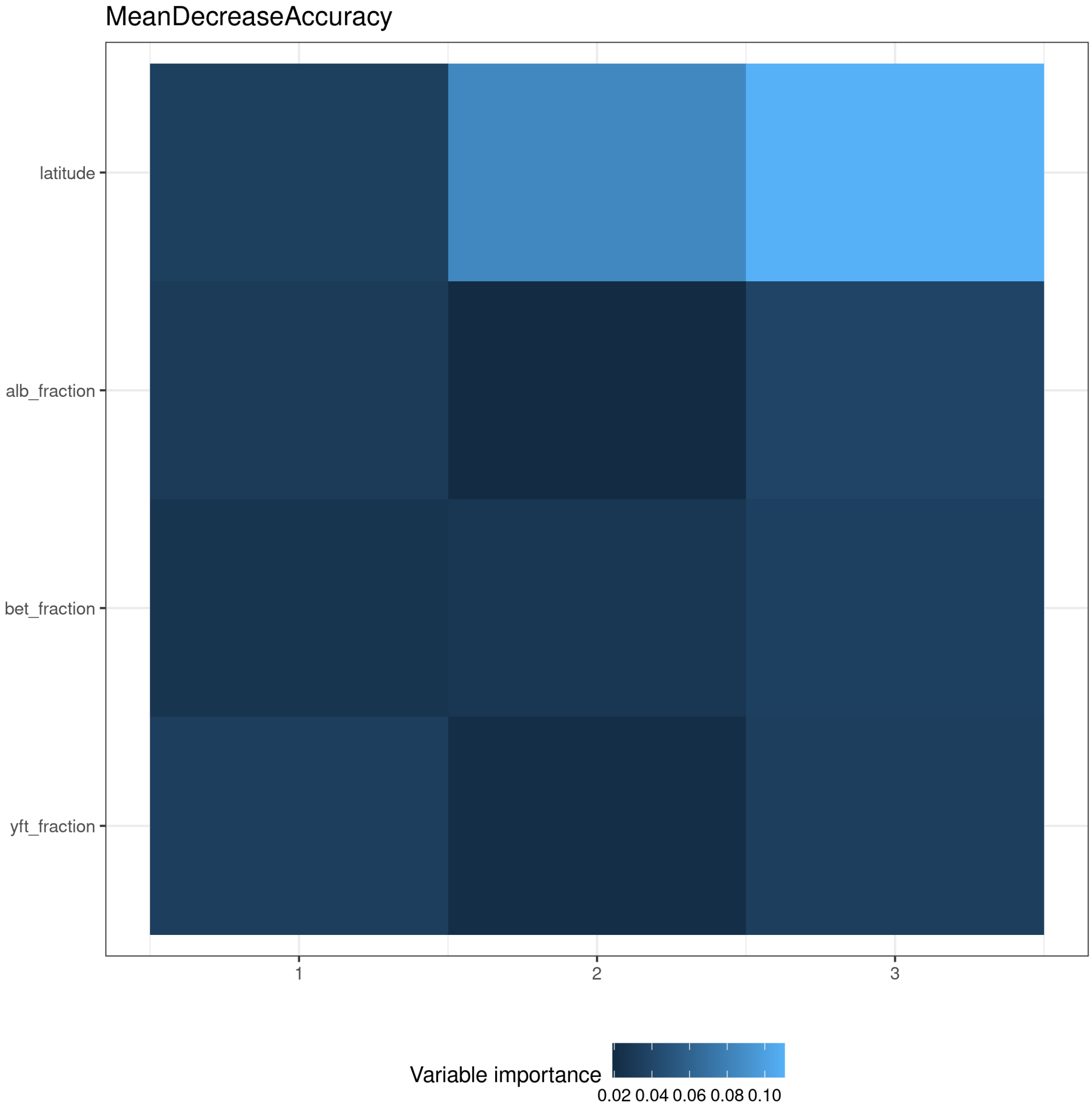

Figure 7: Mean decrease accuracy plot per variables and cluster

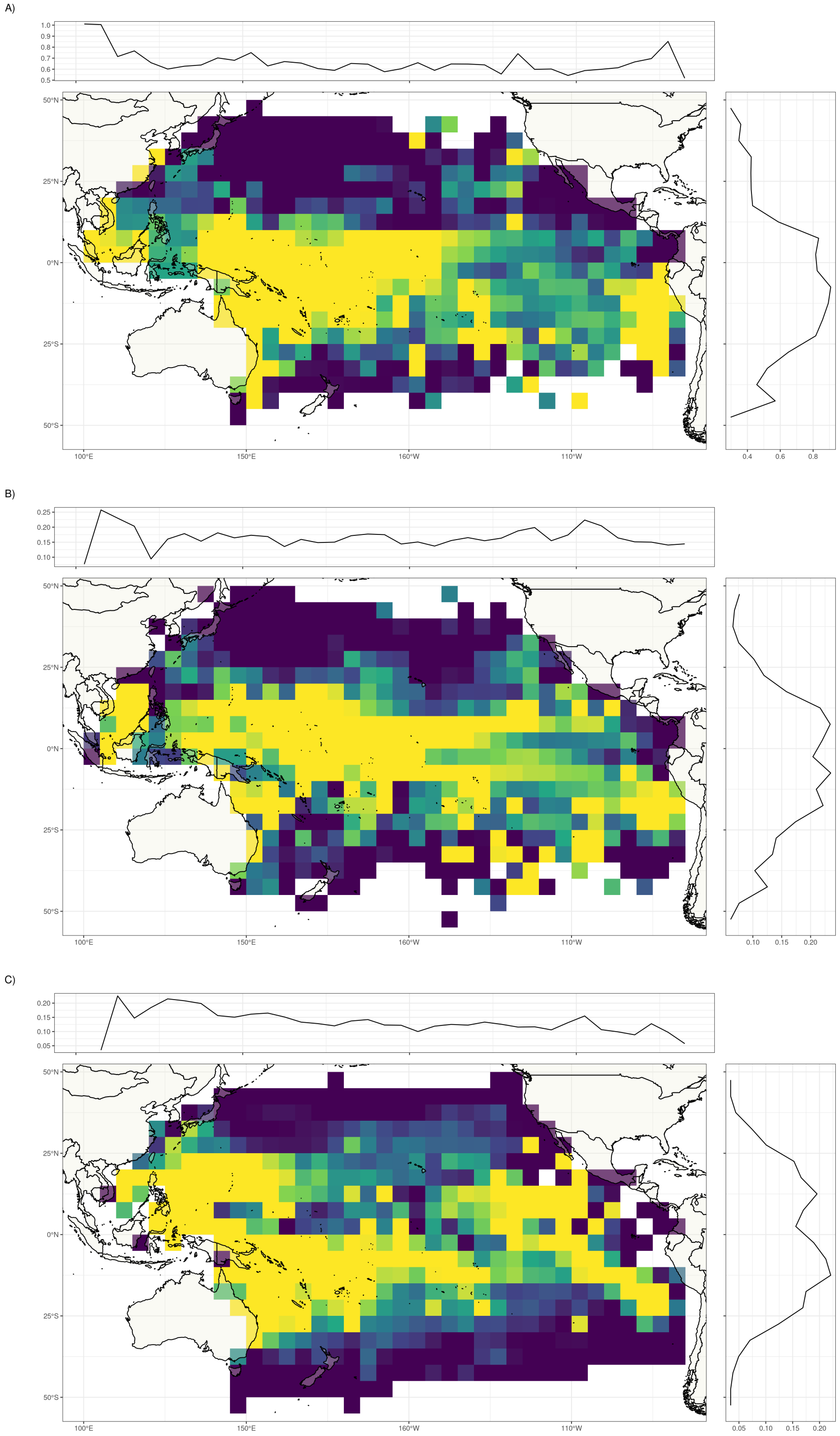

Figure 8: Maps showing the spatial range of the fishery and its average CPUE over time (1960-2022). The curves at the top and right of the maps shows the CPUE average value per longitude and latitude respectively. A through C describe each fishery from 4 to 6 respectively.

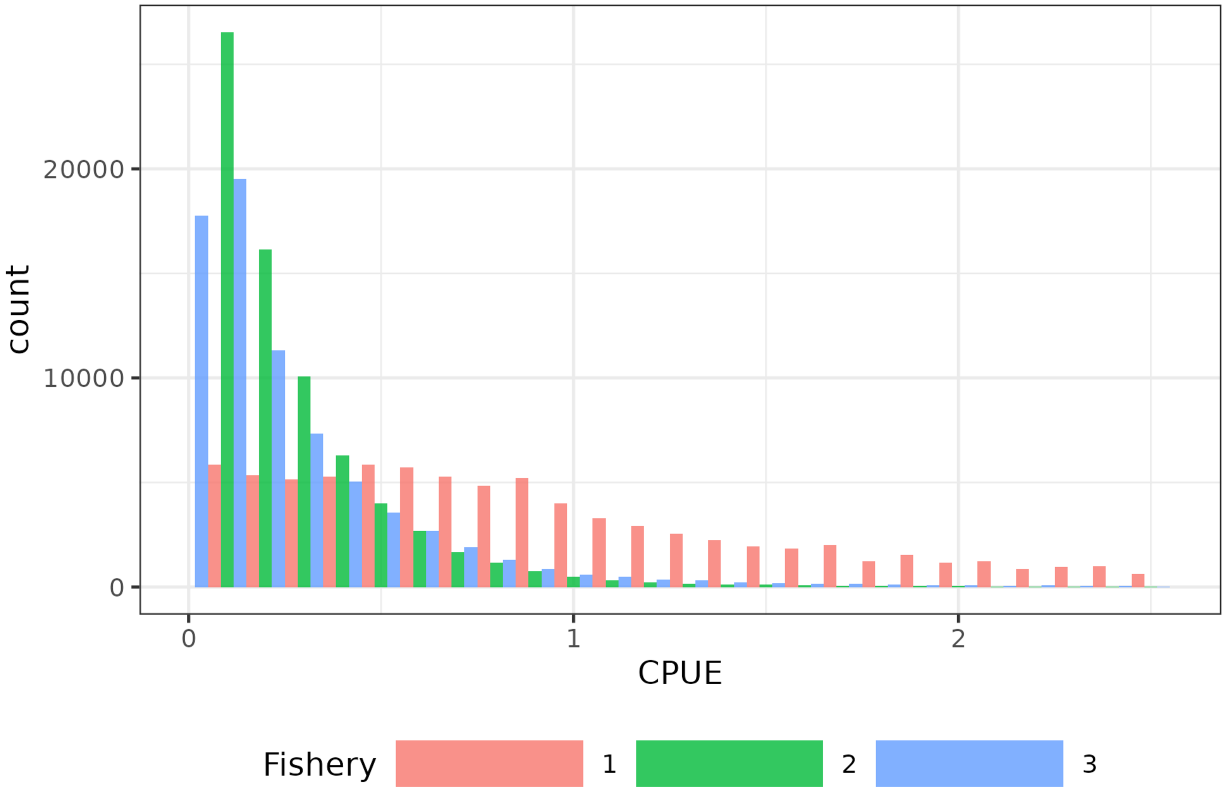

Figure 9: Frequency distribution of CPUE values (yellowfin catch per hundred hooks) across the three delta fisheries.

7 Discussion and Conclusion

Methodological Advances in CPUE Standardization

This study presents a novel approach to partitioning longline fisheries data into homogeneous clusters for integration into the spatially-explicit SEAPODYM model. While modern CPUE standardization for tuna stocks now employs sophisticated spatio-temporal mixed models that effectively address many limitations of earlier generalized linear model approaches [Maunder and Punt, 2004], these statistical methods remain fundamentally pattern-based—modeling correlational relationships between fish distribution and environmental conditions through random effects and covariates.

Our clustering methodology serves a fundamentally different purpose: rather than standardizing CPUE to remove spatial, temporal, and environmental effects, we structure fisheries data to preserve these signals as informative biomass indices for mechanistic modeling. This builds upon previous habitat-based standardization approaches [Bigelow et al., 2002] that addressed similar challenges in estimating relative abundance of tuna species, but extends the concept to support process-based rather than statistical modeling frameworks.

The clustering methodology successfully identified distinct fisheries characterized by relatively consistent catchability and selectivity patterns within the context of environmental variability. Unlike traditional standardization that assumes basin-wide homogeneity and removes the effects of explanatory variables [Maunder and Punt, 2004], our approach recognizes that systematic changes in catchability may reflect meaningful biological processes—such as fish responding to environmental gradients—that should be preserved as model inputs.

This approach addresses a key challenge in mechanistic modeling: establishing linear relationships between CPUE and biomass density at spatial scales appropriate for biological processes. While traditional methods often struggle with year x area interactions [Maunder and Punt, 2004], our methodology explicitly accounts for spatial heterogeneity by establishing fisheries with consistent spatial targeting patterns for integration with spatially-explicit biological processes.

While no fishery maintains perfectly stable catchability over multi-decadal time series—particularly given technological changes including the transition to monofilament and changes in setting depth [Ward and Hindmarsh, 2007]—our methodology identifies fisheries where catchability changes can be attributed to systematic trends rather than structural breaks in fishing practices.

The effectiveness of our approach is demonstrated by the absence of abrupt breakpoints in CPUE and length time series within optimally defined fisheries, indicating that any temporal changes in catchability follow gradual, systematic patterns rather than sudden shifts in fishing behavior. This temporal consistency, while not implying perfect stability, provides the foundation needed for mechanistic models that explicitly account for environmental influences on fish distribution and fishing success.

Our clustering methodology offers complementary advantages to modern statistical spatio-temporal approaches. Rather than competing with these methods, it provides the data structure necessary for process-based modeling where spatial, temporal, and environmental variability represent biological signals rather than statistical noise to be standardized away.

Addressing Key Challenges in Abundance Estimation

A central challenge in using fisheries-dependent data for abundance estimation is the non-random distribution of fishing effort, which tends to concentrate in areas of high fish density. This spatial targeting behavior can lead to hyperstability in abundance indices, where CPUE remains high even as overall abundance declines [Ducharme-Barth et al., 2022].

Our approach addresses spatial targeting through two methodological choices that align with mechanistic modeling requirements. First, by utilizing high-resolution (1°x 1°) gridded data, we establish conditions where relative homogeneity of biomass density can be reasonably assumed within individual grid cells, while capturing broader spatial heterogeneity through the aggregation of cells [Nooteboom et al., 2023]. This fine-scale approach parallels habitat-based standardization efforts [Bigelow et al., 2002] but serves a different purpose: rather than removing spatial effects statistically, we preserve spatial information as biological signals for process-based modeling.

Second, our clustering approach identifies coherent groups of fishing operations characterized by consistent species targeting patterns. While modern spatio-temporal CPUE models effectively handle targeting through covariates and random effects, our methodology structures targeting information differently—grouping fisheries by how they interact with biological processes rather than by statistical properties alone. This distinction becomes important when environmental conditions drive both fish distribution and fishing success, creating signals that statistical models might standardize away but mechanistic models need to preserve.

The issue of effort creep—increasing gear efficiency that can manifest as apparent CPUE increases unrelated to abundance changes [Hamer et al., 2024] was not directly addressed in our data processing methodology. However, SEAPODYM accommodates this phenomenon by allowing catchability parameters to incrementally change through time [Senina et al., 2018], complementing our data-driven approach with model flexibility.

Integration with SEAPODYM and Implications for Biomass Modeling

The fisheries defined through our clustering approach are specifically structured for integration into SEAPODYM, which represents a fundamentally different paradigm compared to traditional stock assessment models. While conventional models typically assume homogeneous biomass distribution within large management areas, SEAPODYM explicitly models spatial heterogeneity and fish movement in response to environmental gradients [Lehodey et al., 2010].

This spatial explicitness offers several advantages for understanding yellowfin tuna population dynamics. At the fine resolution employed in SEAPODYM (1°x 1°), the model can simultaneously account for areas of high fishing mortality without requiring unrealistic recruitment assumptions, as fish movement between adjacent cells provides a mechanistic explanation for local depletion and replenishment patterns. Furthermore, apparent hyperstability in basin-wide CPUE can be mechanistically explained by the non-linear relationship between spatially heterogeneous biomass density and non-randomly distributed fishing effort, without necessarily indicating population decline across the entire Pacific.

The integration of fisheries data with complementary information sources further strengthens SEAPODYM's capacity for robust abundance estimation. Larvae survey data informs recruitment dynamics, while tagging data provides critical information on movement patterns. This multi-source approach compensates for the non-random sampling inherent in fisheries data that tends to bias towards high biomass density areas.

Methodological Limitations and Future Directions

Despite the advances presented in this study, several limitations and areas for future improvement merit discussion. First, our approach to length frequency data encountered significant coverage challenges, with only a small percentage of catch-effort entries having associated length measurements. While we implemented a hierarchical approach to maximize length data utilization, the potential for bias in size selectivity estimation remains. Future work could explore more sophisticated imputation methods or alternative approaches to characterizing size selectivity with sparse data.

Second, the treatment of effort creep through SEAPODYM's time-varying catchability parameter represents a simplification of potentially complex technological and behavioral changes in fishing operations. More explicit modeling of technological transitions or incorporation of vessel-specific characteristics could further refine catchability estimation.

Third, while our clustering approach effectively identified distinct fishery types, the possibility remains that additional unobserved factors influence catchability and selectivity. Sensitivity analyses exploring alternative clustering variables or methodologies could provide insights into the robustness of our fishery definitions. The methodology developed in this study has potential applications beyond SEAPODYM. The principles of identifying homogeneous fishery groups with consistent catchability and selectivity could be adapted for other spatially-structured population models or even conventional stock assessment approaches. For traditional models like MULTIFAN-CL [Fournier et al., 1998], our clustering approach could inform more objective fishery definitions that reduce bias in abundance indices.

Conclusion

This study presents a comprehensive methodology for structuring longline fisheries data into cohesive groups characterized by consistent catchability and selectivity. By combining multiple clustering quality metrics with ecological understanding of tuna biology and fishing operations, we have developed an objective approach to fishery definition that addresses key challenges in using fisheries-dependent data for abundance estimation.

The resulting fishery clusters demonstrate temporal stability in CPUE-abundance relationships while revealing distinct spatial patterns consistent with known habitat preferences of yellowfin tuna. When integrated into SEAPODYM, these carefully defined fisheries enable robust abundance estimation that accounts for spatial heterogeneity and movement dynamics.

As fisheries management increasingly employs spatially-explicit approaches, methodologies that effectively harness the information content of fisheries-dependent data while accounting for its inherent biases become increasingly valuable. The framework presented here represents a step toward more objective, data-driven approaches to fishery definition that can support both scientific understanding and sustainable management of highly migratory tuna populations.

Appendix A Extended data

Table 4: Summary of removed data

Dataset	NA		NULL/Outliers		Mask	
	entries	%	entries	%	entries	%
Captain EC	0	0	5258	0.09	8542	0.15
Observer EC	11554	4.2	2061	0.76	6	0.002
Raised EC	0	0	0	0	689	0.18
Detailed LF	1937382	22.6	336407	3.9	520	0.006
Aggregated LF	304	0.03	0	0	847	0.08

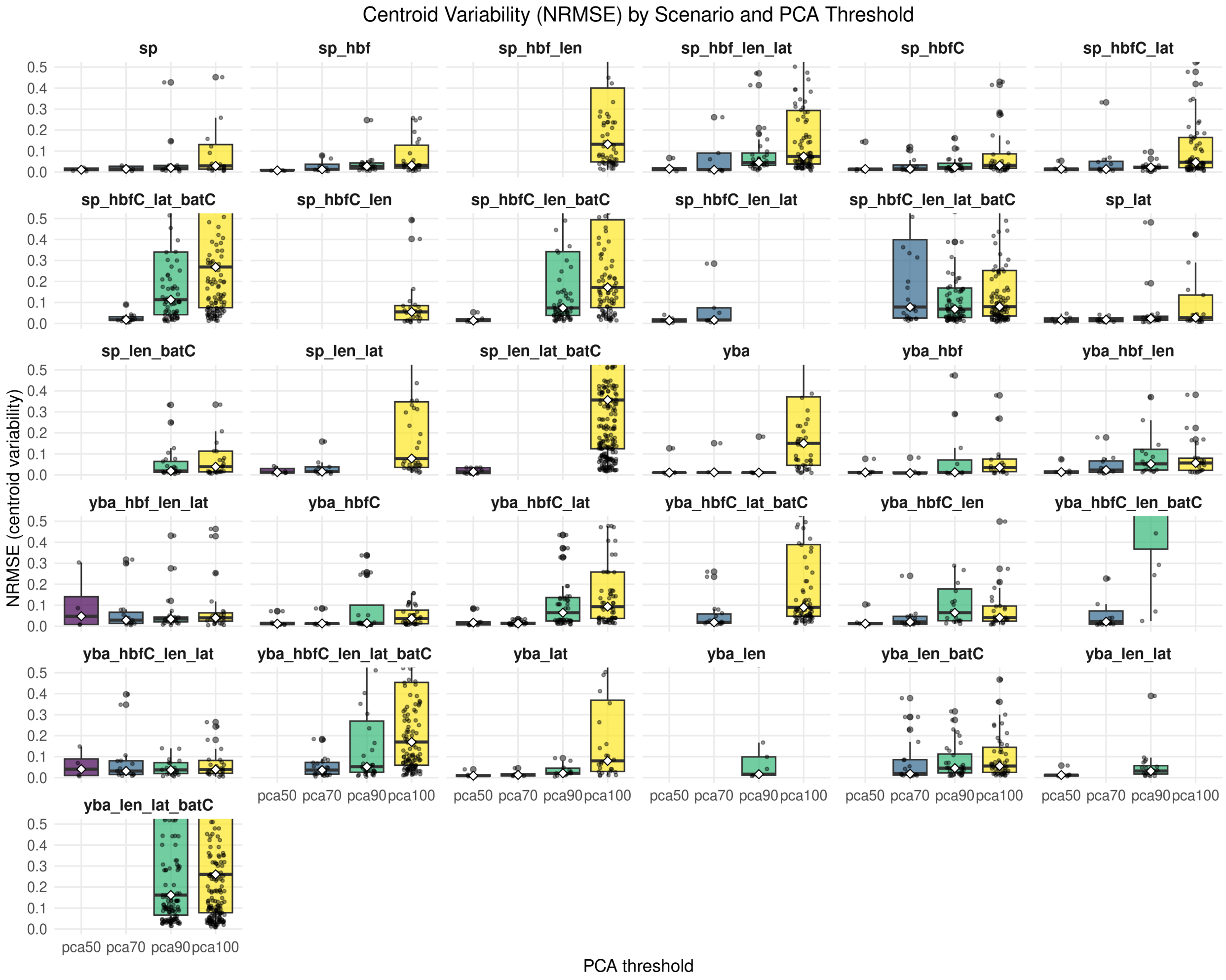

Figure 10: Boxplot showing the normalised root mean square error per scenario.

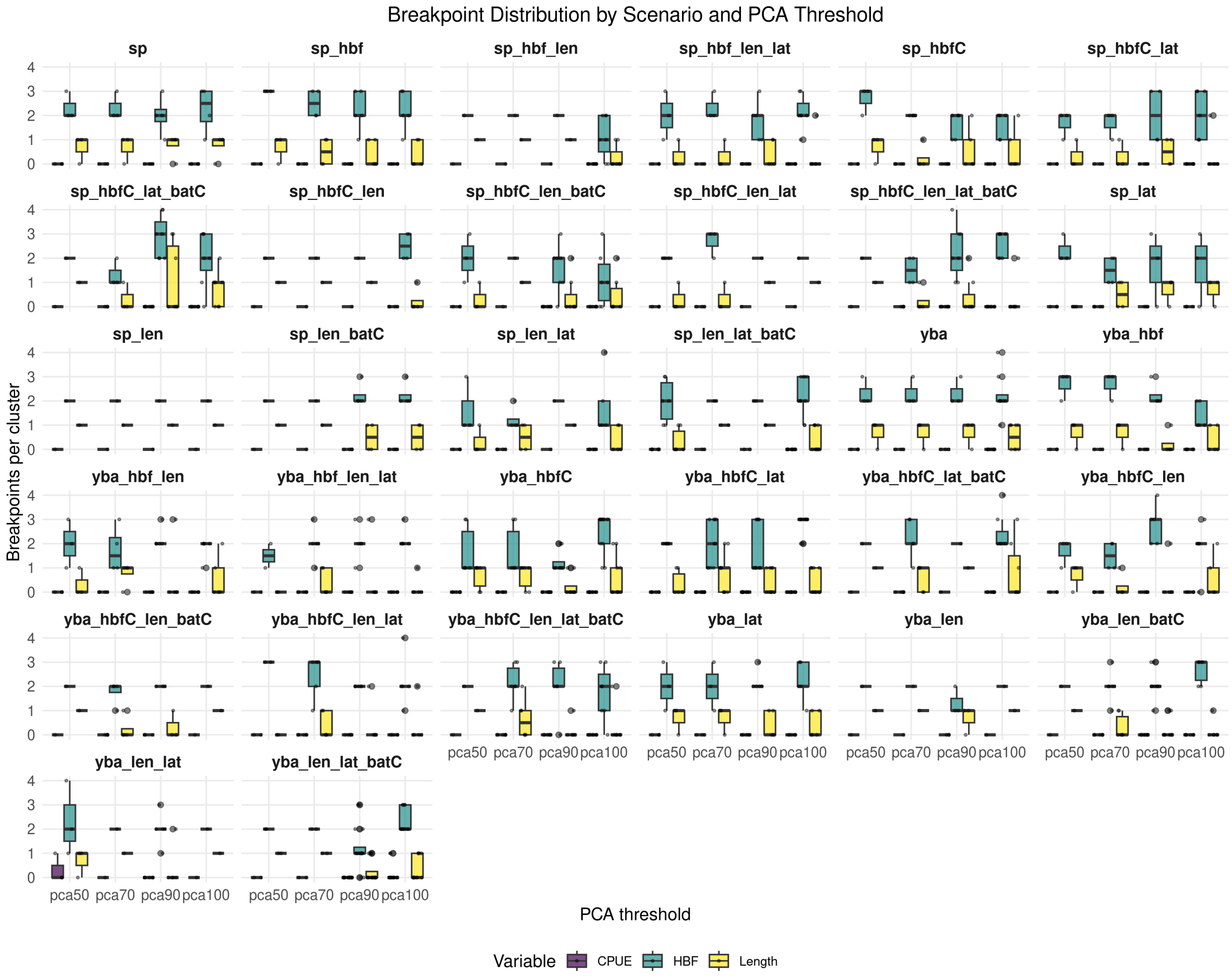

Figure 11: Boxplot showing the number of breakpoint per CPUE, length and HBF per scenario.

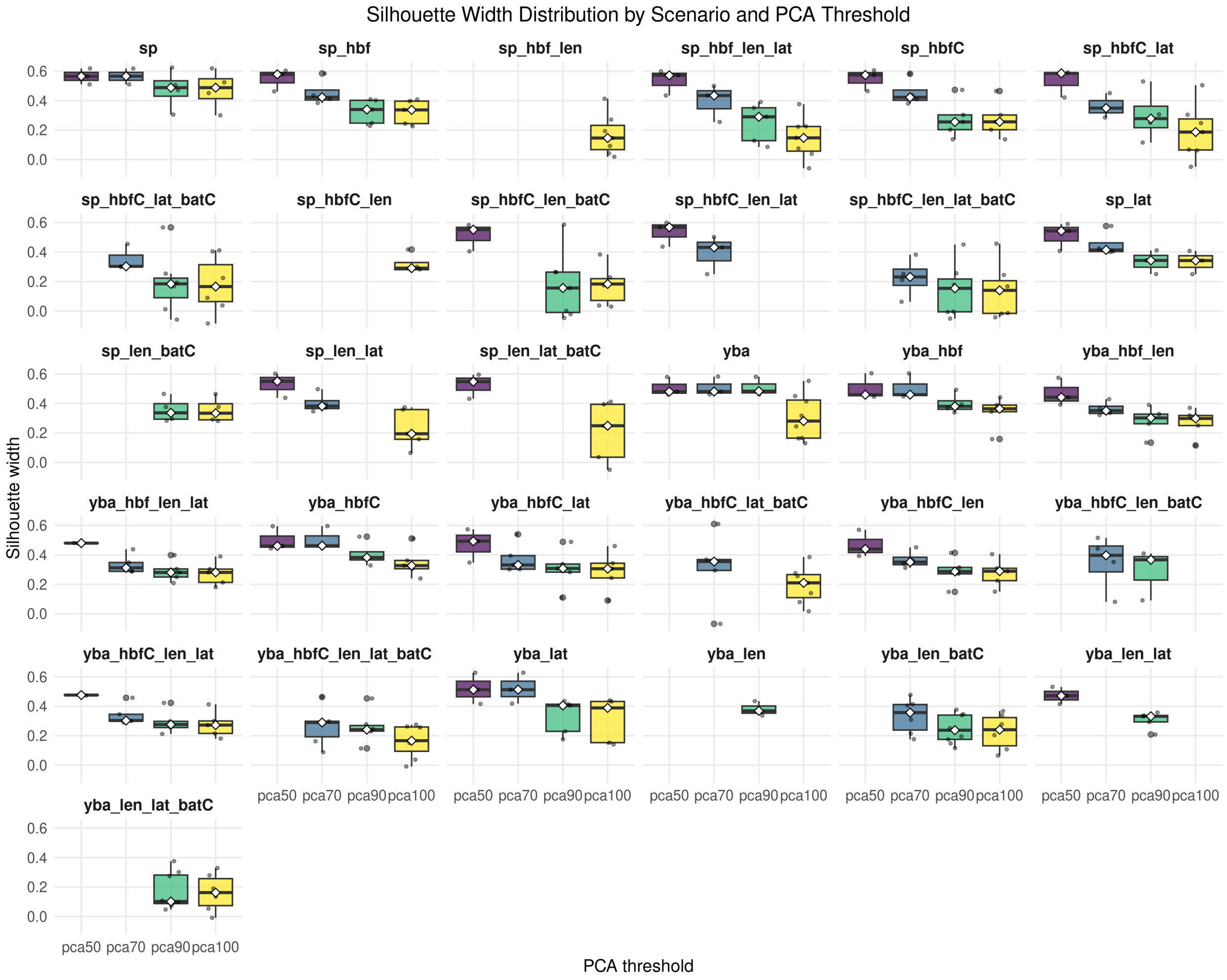

Figure 12: Boxplot showing the average silhouette width per scenario.

8 References

R.J.H. Beverton and S.J. Holt. *On the dynamics of exploited fish populations.* Fisheries Investigations, 1957.

Keith A. Bigelow, John Hampton, and Naozumi Miyabe. Application of a habitat-based model to estimate effective longline fishing effort and relative abundance of pacific bigeye tuna. *Fisheries Oceanography*, 11:143–155, 2002.

J. M. Braccini, M.-P. Etienne, and S. J. D. Martell. Subjective judgement in data subsetting: implications for cpue standardisation and stock assessment. *Marine and Freshwater Research*, 62:734, 2011.

Nicholas D. Ducharme-Barth, Arnaud Gruss, Matthew T. Vincent, Hidetada Kiyofuji, Yoshinori Aoki, Graham Pilling, John Hampton, and James T. Thorson. Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization. *Fisheries Research*, 246:106169, 2022.

Daid A Fournier, John Hampton, and John R Sibert. Multifan-cl: a length-based, age-structured model for fisheries stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, 1998.

Paul Hamer, Inna Senina, Patrick Lehodey, Makoto Nishimoto, Yoshinori Aoki, Naoto Matsubara, and Yuichi Tsuda. Investigating long-term recruitment trends of skipjack tuna in the western and central pacific ocean. *Frontiers in Marine Science*, 2024.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, Andras Horanyi, and Joaquin Munoz-Sabater. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146: 1999–2049, 2020.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *The journal of Educational Psychology*, 1933.

Simon D Hoyle and Mark N Maunder. Standardization of yellowfin and bigeye cpue data from japanese longliners, 1975 - 2004. Technical Report SAR-7-07, Inter-American Tropical Tuna Commission, 2006.

Rebecca Killick, P. Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590–1598, 2012.

Patrick Lehodey, Raghu Murtugudde, and Inna Senina. Bridging the gap from ocean models to population dynamics of large marine predators. *Progress in Oceanography*, 84:69–84, 2010.

Jed Macdonald, Peter Williams, Caroline Sanchez, Emmanuel Schneiter, Suresh Prasad, Marc Ghergariu, Malo Hosken, Aurelien Panizza, Tim Park, Aurelie Guillou, and Simon Nicol. Project 90 better data of fish weights and lengths. Technical Report WCPFC-SC18-2022/ST-IP-04, Oceanic Fisheries Program, 2022.

Mark N. Maunder and Andre E. Punt. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, 70:141–159, 2004.

Peter D Nooteboom, Joe Scutt Phillips, Inna Senina, Erik van Sebille, and Simon Nicol. Individual-based model simulations indicate a non-linear catch equation of drifting fish aggregating device-associated tuna. *ICES Journal of Marine Science*, 80:1746–1757, 2023.

Andre E. Punt. Modelling recruitment in a spatial context: A review of current approaches. *Fisheries Research*, 217:140–155, 2019.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, pages 53–65, 1987.

Inna Senina, John Sibert, and Patrick Lehodey. Parameter estimation for basin-scale ecosystem-linked population models of large pelagic predators. *Progress in Oceanography*, 78:319–335, 2008.

Inna Senina, Patrick Lehodey, Beatriz Calmettes, Morgane Dessert, John Hampton, Neville Smith, Thomas Gorgues, Olivier Aumont, Matthieu Lengaigne, Christophe Menkes, Simon Nicol, and Marion Gehlen. Impact of climate change on tropical pacific tuna and their fisheries in pacific islands waters and high seas areas. Technical Report SC14-EB-WP-01, Western and Central Pacific Fisheries Commission, 2018.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, 63:411–423, 2001.

Stef Van Buuren and Karin Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

Peter Ward and Sheree Hindmarsh. An overview of historical changes in the fishing gear and practices of pelagic longliners. *Reviews in Fish Biology and Fisheries*, 17:501–516, 2007.

Jinn-Shing Weng, Ming-An Lee, Kwang-Ming Liu, Hsing-Han Huang, and Long-Jing Wu. Habitat and behaviour of adult yellowfin tuna in the waters off southwestern taiwan. *Aquatic Living Resources*, 30:34, 2017.

Juanying Xie and Shuai Jiang. A simple and fast algorithm for global k-means clustering. In *2010 Second International Workshop on Education Technology and Computer Science*, pages 36–40, 2010.