DESIGN DOCUMENT FOR THE NORTH ATLANTIC SWORDFISH MANAGEMENT STRATEGY EVALUATION. OPERATING MODEL (OM) AND OBSERVATION ERROR MODEL

L. T. Kell[1], P. Levontin[1]

SUMMARY

This document outlines the potential specifications of an Operating Model (OM) and an Observation Error Model (OEM) that could be used in initial simulation trials to evaluate alternative management strategies for North Atlantic swordfish. The OM that can be conditioned on a variety of data sets and hypotheses while the OEM can be used to evaluate different data collection regimes e.g. aerial surveys, tagging programmes, catch and catch per unit effort (CPUE) and size to age conversions.

RÉSUMÉ

Le présent document décrit les spécifications potentielles d'un modèle opérationnel (OM) et d'un modèle d'erreur d'observation (OEM) pouvant être utilisés dans des essais de simulation initiaux afin d'évaluer des stratégies de gestion alternatives pour l'espadon de l'Atlantique Nord. L'OM peut être conditionné sur une gamme de jeux de données et d'hypothèses tandis que l'OEM peut être utilisé pour évaluer les différents systèmes de collecte de données tels que les prospections aériennes, les programmes de marquage, les captures et les captures par unité d'effort (CPUE) et les conversions de taille en âge.

RESUMEN

Este documento describe las especificaciones potenciales de un modelo operativo (OM) y un modelo de error de observación (OEM) que podría utilizarse en ensayos de simulación iniciales para evaluar estrategias de ordenación alternativas para el pez espada del Atlántico norte. El OM puede condicionarse en una variedad de conjuntos de datos e hipótesis mientras que el OEM puede utilizarse para evaluar diferentes regímenes de recopilación de datos, por ejemplo: prospecciones aéreas, programas de marcado, captura y captura por unidad de esfuerzo (CPUE) y conversiones talla a edad.

KEYWORDS

Atlantic, Swordfish, Management Strategy Evaluation,
Operating Model, Observation Error Model

[1] Centre for Environmental Policy, Imperial College London, London, United Kingdom.

Relevant Sections of the Agreed Work Plan

This is a draft outlining the specifications of the Operating Model (OM) and the Observation Error Model (OEM) that will be used in the initial runs to evaluate alternative management strategies for North Atlantic swordfish.

The Relevant sections of the agreed work plan are:

i) OM that can be conditioned on a variety of data sets and hypotheses (using as a base the last N-SWO stock assessment: Stock Synthesis 3)
ii) An OEM that can be used to evaluate different data collection regimes e.g. aerial surveys, tagging programmes, catch and catch per unit effort (CPUE) and size to age conversions.

This paper is a deliverable related to Milestones 1, 5 and 6 with the corresponding deadlines for respective stages of the Design Document that:

The first milestone is that by the 1st of September an outline of the OM and Observation Error Model will be delivered to the WG. This is that document.

Introduction

This document details the design of the Operating Model (OM), including the Observation Error Model (OEM) for North Atlantic swordfish. Both the OM and OEM are implemented in R using FLR (Kell *et al.*, 2007) which uses S4 Classes for object oriented programing. The FLR project has been developing and providing fishery scientists with a powerful and flexible platform for quantitative fisheries science based on the R statistical language. The guiding principles of FLR are **openness**, through community involvement and the open source ethos, **flexibility**, through a design that does not constraint the user to a given paradigm, and **extendibility**, by the provision of tools that are ready to be personalized and adapted. The main aim is to generalize the use of good quality, open source, flexible software in all areas of quantitative fisheries research and management advice.

The OM can be conditioned on a variety of data sets and hypotheses, however, in the first instance conditioning is based on the last N-SWO stock assessment conducted using Stock Syntheses (Methot and Wetzel, 2013). The OM is used to describe resource dynamics in simulation trials and the Observation Error Model (OEM) simulates pseudo data for stock assessments that are part of management procedures being tested. Simulated data sets can take the form of: catch data, catch per unit effort (CPUE), length compositions, or different potential data collection regimes such as aerial surveys, tagging programmes and size to age conversions.

Most of the OMs so far developed by the tuna Regional Fishery Management Organisations (tRFMOs) were developed using on a stock assessment paradigm based on the current stock assessment. Although the current approach of using assessment models as the basis for OM design is a reasonable starting point, improvements need be made to the adequacy of conditioning the OMs, especially with respect to enabling the inclusion of other important processes and uncertainties that difficult to account for in stock assessments but affect the robustness of advice (Sharma *et al.*, submitted).

Model Validation Procedure and the Tuna Regional Fisheries Management Organisations

As agreed in the work plan a procedure for model validation will be developed and described in the design document, i.e.

Step by Step Summary of a Procedure for Selecting Oms

1. Create factorial design on base case based on an agreed set of uncertainties

2. For main effects check that

   a. Model has converged and results are plausible
   b. Model diagnostics
   c. Dimensionality reduction

3. Propose main effects

4. Run full grid interactions and repeat step 2

5. Propose Reference Set for OM

6. Propose Robustness Set for OM

The work plan also notes that this is a topic that is on the agenda of the tRFMO MSE WG, and it was agreed to collaborate with that group and provide feedback on the recommendations made and to follow their advice in developing North Atlantic swordfish MSE.

The tRFMO MSE WG made several recommendations in 2018 on the Conditioning of OMs, these included:

1. With respect to OMs, the Group **advises** that it is valuable to limit their number to that needed to adequately address the key uncertainties, with a focus on those that may have management implications in the future. However, it **stresses** that this limitation should not be taken too far – the OMs should consider a range of plausible scenarios which is sufficiently broad that tested MPs or HCRs do not require amendment or retesting too often.

2. The Group also **stresses** that it **essential** that all OMs are adequately conditioned i.e. ensure that they are sufficiently consistent with the historical data to be considered plausible. Whilst conditioning is a case-specific process, there are some general guidelines that should be followed including: the use of standard model fit diagnostics for indications of model misspecification (automated where possible); focusing on the conditioning of 'limit' cases and which may be sufficient to justify the assumption that conditioning in between these is adequate.

Conditioning an OM on a full factorial assessment grid may not be possible or advisable as not all interactions may converge or they may produce implausible parameter estimates (see Kell and Mosqueira, 2017). Factors in the grid are not independent, therefore some combinations are likely to be highly improbable and not supported by the data. An agreed procedure for model validation is therefore needed in order to filter OMs in order to retain those that are plausible and interesting and to discard others.

Therefore as agreed in the work plan this document is a platform to agree on a procedure to identify a set of reference and robustness trials with the Swordfish working group. Terminology is based on that of Rademeyer *et al.*, (2007).

The OM reference set is a small set of OMs, which include the most important uncertainties in the model structure, parameters, and data - alternative scenarios which are believed to have high plausibility and to cause major impacts on performance statistics of management procedures. In complement, robustness tests examine the performance of an MP across a fuller range of plausible scenarios, including a wider than the reference set of OMs. Although robustness is required across a range of uncertainties there is no need to have robustness to OMs that are implausible.

It is important to explicitly address each of the major sources of uncertainty, or at least indicate how the uncertainties reflected were selected. Punt *et al.* (2016) states that as best practice, minimally, an MSE should consider (i) process uncertainty, in particular, variation in recruitment about the stock - recruitment relationship; (ii) parameter uncertainty relating to (a) productivity and (b) the overall size of the resource; and (iii) observation error in the data used when applying the management strategy, required to model the OEM. It is also crucial that the conditioning of the OMs is adequate to ensure that there is no evidence of systematic misfits to data, unless a particular data source itself is highly uncertain.

Feedback also relaxes the requirement of having to have an exact model of the system being managed (Punt *et al.* 2016). However, not accounting for some critical aspect of uncertainty, like non-stationarity of biological parameters (such as, growth or natural mortality) can lead to a false sense of security.

North Atlantic Swordfish Case Study

The Swordfish working group, discussed different options for conditioning the OM taking into consideration a variety of uncertainties. It was decided that the initial approach will be to use Stock Synthesis (SS) for OM conditioning, by either using the uncertainties identified during the stock assessment or by considering further sources of uncertainty that were not identified at the time but that the goodness of fit diagnostics as being important, i.e. due to data conflicts and model misspecification identified using a base case or sensitivity runs, or the group considers important to be incorporated into the OMs that describe the resource dynamics.

The main grid of uncertainties discussed by the WG are summarised in **Table 1** while a larger set of uncertainties is shown in **Table 2**. These cover a range of scenarios related to model structure, fixed values for difficult to estimate parameters, and alternative data weightings and datasets. If the entire grid (i.e. a full factorial design) is run just on the main grid (**Table 1**) this will require 28,800 SS assessments to be conducted. Some factors, however, represent alternative hypotheses for particular series of observations and may therefore be mutually exclusive. For example a lack of older fish in catches could be explained by either high natural mortality or dome shaped selection pattern but not both as seen in the Indian Ocean Albacore MSE (Kell and Mosqueira, 2016).

As the grid is being used to represent uncertainty arising from conflicting or unreliable data sources, it is likely that the resulting OM will show clear evidence of data conflicts and model misspecification (Maunder and Piner 2014, 2107 and Maunder and Punt, 2013). Furthermore many will not converge or produce plausible representations of the stock (either in terms of parameter values or historical trajectories). While, initial fits will suggest alternative assessment hypotheses that should be explored.

This is the stage where the procedure for filtering OMs will be applied. Some of the more commonly used methods for selecting models, such as AIC, cannot be used in the context that involves different datasets, for instance where OMs are based on alternative CPUE series or conditioned on alternative effective sample sizes for length frequencies (Kell et al. 2016). A related issue is how to weight the scenarios for which OMs are developed in relation to their relatively plausibility. The tRFMO MSE WG agreed that this is an important and difficult issue that should be taken up with high priority in future meetings.

If only the main effects are run then 22 SS assessments would need to be conducted, which will require a considerable amount of work in terms of development, validation and checking for convergence, data conflicts model mis-specification.

It is likely that various combinations of scenarios in **Table 1** may not be consistent with the data i.e. not considered plausible. For example, some combinations of selectivity, M and steepness may give estimates of carrying capacity and population growth rate that are not consistent with the biology. In the case of the Indian Ocean Albacore MSE which used a full factorial design many runs were found to be implausible (Kell and Mosqueira, 2017).

As in most stock assessments there may be patterns in the residuals that indicate model mis-specification. For example patterns in the residuals, in such instances data series exhibiting patterns cannot be used to simulate data for use in the MP, since in the case of a CPUE series they will not provide a reliable index of abundance. The selection of hypotheses for the OM will therefore need to consider scenarios corresponding to relative data weightings, of fleets, CPUE series and length composition. These are not always possible to pre specify as alternative weighting are normally suggested by data conflicts after running a range of sensitivity runs and inspecting appropriate diagnostics.

Some of the identified uncertainties present exceptional challenges to conditioning on data and for MSE modelling, in particular: sexual segregation, stock structure and environmental scenarios. These are issues that as of yet have not been incorporated into stock assessment advice by the SCRS.

One of the options for incorporating uncertainty related to future environmental changes is to add scenarios characterised by the believed implications of future environmental scenarios on the parameters that drive population dynamics, such as M, recruitment variability, steepness etc., see Punt et al., (2014) for a discussion of the strengths and limitations of the various approaches. These are potential scenarios for robustness trials.

Methods

Following the recommendations of the tRFMO MSE WG that OMs

> *"should consider a range of plausible scenarios which is sufficiently broad that tested MPs or HCRs do not require amendment or retesting too often"*

We therefore propose a procedure for selecting OM scenarios that are a) plausible; and b) likely to provide a different perspective on alternative MPs.

The logic for selecting OMs is based on the recommendations of the tRFMO MSE WG, namely that it is valuable to limit the number of OMs to that needed to adequately address the key uncertainties, with a focus on those that may have management implications in the future. If a combination of assumptions produces simulations that are wildly inconsistent with observations then these combinations should not be used in testing management procedures; and if two OMs are likely to say exactly the same things about all of the MPs in question then it is not necessary to run them both.

Procedure for conditioning the OM on data and knowledge, and for validating, rejecting and weighting hypotheses

Although a variety of data types can be included in integrated stock assessments to simultaneously provide information on all estimated parameters, conflicts between data can affect the estimates of important parameters rendering the OM model conditioned on such assessments biased and possibly misleading representations of the resource. Conditioning the OM on an integrated assessment is therefore an iterative process. Fits for each scenario will be examined and a range of diagnostics tests used to detect misspecification and data conflicts. We propose a procedure, that will be used first to condition the OMs on the main effects, to set bounds on parameter values and evaluate potential data conflicts; after which the important interactions will be proposed and investigated, using the following steps (detailed examples are being developed and are being placed on github), i.e. 1. Check convergence and plausibility of estimates; 2. Use standard model fit diagnostics for indications of data conflicts and model misspecification; and ***3*** Dimensionality reduction, i.e. as stated above if two OMs are likely to say exactly the same things about all of the MPs in question then it is not necessary to run them both.

The tRFMO MSE WG recommended that initially focus should be on the conditioning of 'limit' cases, i.e. the extreme levels of a factor, since this should be sufficient to justify the assumption that conditioning in between these is adequate, i.e. if a model does not fit the data for an upper or lower value of a fixed parameter (such as M or steepness) then there is no point in running any interactions.

Therefore the first step will be to run the main effects to i check convergence and plausibility of estimates in order to determine the appropriate limits to factor levels; then use standard model diagnostics to help identify data conflicts that may require alternative data weighting scenarios, e.g. down weighting length composition data compared to indices of abundance or apply different weights to CPUE series; and then identify potential model mis-specification that may require additional scenarios. Following this aa set of factors and second level interactions will be proposed and steps 1 and 2 of the procedure repeated.

Step A: Convergence Tests and Plausibility of estimates

The Indian Albacore MSE, where the OM was conditioned using SS (Kell and Mosqueira, 2017), and the study of Carruthers et al. (2017) showed that factors that have the biggest impact on estimates from stock assessment are often those for which there is little information in the data to fit, i.e. steepness of the stock recruitment relationship, natural mortality and the shape of the selection function. In addition when conditioning an OM on a factorial design a number of runs may lead to unrealistic estimates of quantities such as virgin biomass (B_0), population growth rate (r), or trajectories of the historical stock. Such runs can be filtered out based on the upper and/or lower limit of the estimates of B_0 and r. For example an upper limit was estimated based on the relationship between carrying capacity (K) and suitable habitat for all global albacore stocks (Kell and Mosqueira, 2017) and a similar procedure was developed for swordfish (Arrizabalaga *et al.,* 2017). While life history theory can be used to estimate r.

Step B. Model Diagnostics

Residuals from stock assessment fits include both process and measurement error and pattern of residuals may indicate serial correlation in sampling/observation error or model misspecification. It is particularly important to identify model misspecification. In a simulation exercise Carvalho *et al.* (2017) showed that residual analyses were easily the best detector of misspecification of the observation model. Therefore tests will be used to evaluate goodness of fit (e.g. SEDAR 40, 2015), for example runs tests, to compare residuals to fits across datasets. If the process of interest shows only random variation, the data points will be randomly distributed around the median. Non-random variation may present itself in several ways. If the process centre is shifting due to improvement or degradation unusually long runs of consecutive data points may be seen on the same side of the median or the graph crosses the median unusually few times. Violation of randomness may be due to data conflicts and/or model misspecification and also means that the OEM cannot be developed based on the datasets that fail the test.

This will be an iterative process as in any assessment, since problems due to model misspecification will require alternative hypotheses to be tested as conditioning scenarios.

The OEM which generates fishery-dependent and/or fishery-independent resource monitoring data also needs to be specified. This should be based on the residuals from the assessment, this will require, however, that the residuals from the fits to the indices of abundance pass appropriate goodness of fit diagnostics. The most common way to determine a model's goodness-of-fit is presented in Cox and Snell (1968). Residuals are examined for patterns to evaluate whether the model assumptions have been met (Wang *et al.*, 2009).

Step C. Dimension Reduction

If two OMs are likely to say exactly the same things about all of the MPs in question then it is not necessary to run them both. Therefore as conducted for Indian Ocean Albacore and Atlantic bigeye and yellowfin a dimension reduction technique will be used (e.g. PCA, clustering, LASSO regression) to identify those combinations of scenarios in the grid that have a significant effect on the model variance. This will help in selecting a tractabultsle number of OMs for the reference and robustness sets.

Examples and Documentation

Examples will be placed on the github site with code in Rmd and figures for illustration, detailing the procedure. FLR has already been used extensively to perform stock assessment and conduct MSE, there are therefore a wide variety of material available on goodness of fit diagnostics and condition of OMs, see https://github.com/iotcwpm, and North Atlantic albacore work at https://github.com/iccat, https://github/lauriekell/mydas, the tRFMO MSE WG report, and associated papers at papers at https://github.com/laurieKell/xval/wiki

Results

Here we present a folio of results related to:

I. Summaries the 2017 Assessment results including derived quantities.
II. Indices of abundance
III. Residual Analysis and Observation Error Model
IV. Runs plots
V. Parsimonious Grid
VI. Time Series

The intention is to provide material for discussion rather than to over interpret the results

Assessment Results

Figure 1 shows the Kobe phase plot for North Atlantic swordfish from 2017 showing the parameter uncertainty in the estimates of B/BMSY and F/FMSY in the two assessments methods (SS and BSP) used to provide advice. The stock assessment by assessment method. **Figure 2** combines the two assessments and presents with marginal distributions.

Figure 3 contrasts parameter uncertainty from a single run using MCMC simulations, with parameter uncertainty. This shows that statistical uncertainty estimated in a single model run is greater than uncertainty due to value uncertainty related to fixed parameters.

The time series of recruitment, SSB, catch and fishing mortality and summarised in **Figure 4.**

Figure 5, **6, 7, 8** and **9** summarise the length distributions by age from the age length key (ALK) generated by SS3 for scenarios **base, h.6, h.75**, **m.1**, **m.3**.

Figure 10 summarised the stock recruitment relationship fitted by SS, and **Figure 11** shows the cross Spearman correlations between SSB and recruitment, a positive lag indicates that a stock-recruit relationship and a negative lag a recruit-stock relationship. While **Figures 12, 13, 14, 15, 16** show the respective goodness of fit diagnostics.

The equilibrium curves with reference points are shown for each scenario in **Figures 17, 18, 19, 20**, **21.**

Indices of abundance

This section looks at the catch per unit effort (CPUE) data in order to identify conflicts in the data and to help hypotheses that may help setting up stock assessment scenarios to run.

The CPUE time series are plotted in **Figure 22**, to help compare trends by stock a lowess smoother is fitted to year using a general additive model (GAM).

To look at potential conflicts in the data, i.e. deviations from the overall trends, the residuals from the fits are compared in **Figure 23**. Conflicts between indices can be identified by looking for patterns in the residuals.

If indices represent the same stock components then it is reasonable to expect them to be correlated, if indices are not correlated or negatively correlated, i.e. they show conflicting trends, this may result in poor fits to the data and bias in the estimates. Therefore the correlation between the indices is evaluated in **Figure 24**, the lower triangle show the pairwise scatter plots between the indices with a regression line, the upper triangle the correlation coefficients and the diagonal the range of observations. A single influential point may cause a strong spurious correlation there it is important to look at trends as well as the correlation coefficients, since as a strong correlation could be found by chance if a few key points coincide.

The correlations can be used to select groups that represent a common hypotheses about the evolution of the stock, therefore **Figure 25** shows the results from a hierarchical cluster analysis using a set of dissimilarities.
Correlations between series may be lagged due to indices being dominated by particular age classes so the cross-correlations (lagged by -10 to 10 years) are plotted in **Figure 26**. The diagonals show the autocorrelations as an index is lagged against itself.

Figure 27 shows the selection patterns of the various indices used to calculate the indices, to help in interpreting any cross correlations

Residual Analysis and Observation Error Model

Analysis of residuals is perhaps the most common way to determine a model's goodness-of-fit (Cox, 1968). Residuals are examined for patterns to evaluate whether the model assumptions have been met. Many statistics exist to evaluate the residuals for desirable properties. A non-random pattern of residuals may indicate that some heteroscedasticity is present, or there is some leftover serial correlation (serial correlation in sampling/observation error or model misspecification). Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test.

Departures from the assumption that the index is proportional to the stock can also be seen by plotting the residuals by time (**Figure 28)**, there do appear to be patterns in the residuals that may suggest conflicts between the data series. **Figure 29** plots the observed CPUE against the fitted (the blue line is a linear regression fitted to the points and the black line is y=x) as the index is assumed to be proportional to abundance the points should fall either side of the $y=x$ line.

Autocorrelated residuals within indices may be due to year-class effects and allow the importance of factors not included in the standardisation of the CPUE to be identified. Autocorrelation may mean that the estimated parameters are biased, autocorrelation can be checked by plotting the residuals against each other with a lag of 1 (**Figure 30**), if there is no autocorrelation then the fit to the points (blue line) will be horizontal. The error distribution was checked by plotting the observed and the predicted quantiles for a given distribution e.g. for the normal distribution (**Figure 31**).

The variance of the distribution can be checked by plotting the residuals against the fitted values (**Figure 32**).

Runs plot

If the process of interest shows only random variation, the data points will be randomly distributed around the median. Random meaning that we cannot know if the next data point will fall above or below the median, but that the probability of each event is 50%, and that the data points are independent. Independence means that the position of one data point does not influence the position of the next data point, that is, data are not auto-correlated.

If the process shifts, these conditions are no longer true and patterns of non-random variation may be detected by statistical tests. Non-random variation may present itself in several ways. If the process centre is shifting due to improvement or degradation we may observe unusually long runs of consecutive data points on the same side of the median or that the graph crosses the median unusually few times. The length of the longest run and the number of crossings in a random process are predictable within limits and depend on the total number of data points in the run chart (Anhoej, 2014).

A shift signal is present if any run of consecutive data points on the same side of the median is longer than the prediction limit, round(log2(n) + 3). Data points that fall on the median do not count, they do neither break nor contribute to the run (Schilling, 2012).

A crossings signal is present if the number of times the graph crosses the median is smaller than the prediction limit, qbinom(0.05, n - 1, 0.5) (Chen, 2010). n is the number of useful data points, that is, data points that do not fall on the median. The shift and the crossings signals are based on a false positive signal rate around 5% and have proven useful in practice.

Figure 33 show runs charts for base scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings, results from the other scenarios are shown in **Figures 34, 35, 36, 37.**

Parsimonious Grid

In many of the tRFMO MSE Operating Models were developed as factorial designs, where scenarios corresponded to a number of factors with levels. For example difficult to estimate parameters such as steepness and natural mortality were fixed at a range of values and then all the possible combinations run. All the combinations are referred to as the Operating Model grid.

Punt et al. (2014 state that as best practice, minimally, a MSE should consider (i) process uncertainty, in particular, variation in recruitment about the stock – recruitment relationship; (ii) parameter uncertainty relating to (a) productivity and (b) the overall size of the resource; and (iii) observation error in the data used when applying the management strategy. It is also crucial that the conditioning of the OMs is adequate to ensure that there is no evidence of systematic misfits to data.

To explore the variability in the grids a variety of summary statistics were calculated using the Operating Model grids for East Atlantic and Mediterranean bluefin, Indian Ocean albacore and swordfish, and North Atlantic Ocean albacore and swordfish. Summary statistics include MSY reference points, the production function and current stock status.

So that the main features in the grids could be summarised the correlations between the statistics were then estimated and used to order the statistics into similar groups. As well as using the raw correlations the absolute value of the correlation coefficient were used to group the variables to show the strength of the correlations.

Figure 38 show the correlation matrix for Albacore Indian Ocean, while for ease of comparison **Figure 39** the correlation matrix based on absolute values. **Figures 40** and **41** show the same figures for North Atlantic Ocean Albacore and **Figure 42** and **43** for Indian Ocean swordfish.

These correlation matrices are then used to cluster the grid scenarios, **Figures 44, 45** and **46** show the cluster selection for Albacore in the Indian Ocean; **Figures 47, 48** and **49** show the cluster selection for Albacore in the North Atlantic Ocean; and **Figures 50, 51** and **52** show the cluster selection for swordfish in the Indian Ocean.

Time Series

Time series of SSB and recruitment are plotted to better understand the dynamics, i.e. variation over time in **Figures 53** and **54** for Albacore in the Indian Ocean, **Figure 55** and **56** for Albacore in the North Atlantic Ocean; and **Figure 57** and **58** for swordfish in the Indian Ocean recruits.

Discussion and Conclusions

The intention of the paper is to present examples of the type of summary statistics that can be used to compare MPs and so we do not wish to over interpret them.

References

Anhoej, J. (2015). Diagnostic Value of Run Chart Analysis: Using Likelihood Ratios to Compare Run Chart Rules on Simulated Data Series. PLoS ONE 10(3): e0121349.

Arrizabalaga H, Kell L., and Coelho R, SCRS/2017/073.A first approximation to relative habitat size for swordfish stocks.

Chen, Z. (2010). A note on the runs test. Model Assisted Statistics and Applications 5, 73-77

Kell, L.T., Kimoto, A. and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries research, 183, pp.119-127.

Maunder, M. N. and Piner, K. R. 2014. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science, 72(1):7–18, 2014.

M. N. Maunder and K. R. Piner. 2107. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fisheries Research, 192: 16–27.

Maunder, M.N. and Punt, A.E. 2103. A review of integrated analysis in fisheries stock assessment. Fisheries Research, 142:61–74.

Methot, R. D., and Wetzel, C. 2013. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142: 86–99.

Punt, A.E., A'mar, T., Bond, N.A., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A., Haltuch, M.A., Hollowed, A.B. and Szuwalski, C., 2013. Fisheries management under climate and environmental uncertainty: control rules and performance simulation. ICES Journal of Marine Science, 71(8), pp.2208-2220.

Punt, A. E., Butterworth, D. S., de Moor C.L., De Olivera, J. A.A., and Haddon, M. 2016. Management strategy evaluation: Best Practices. Fish & Fisheries 17: 303-334.

Carruthers, T., Kell, L. and Palma, C., 2017. Accounting for uncertainty due to data processing in virtual population analysis using Bayesian multiple imputation. Canadian Journal of Fisheries and Aquatic Sciences, 75(6), pp.883-896.

Carvalho, F., Punt, A.E., Chang, Y.-J., Maunder, M.N., and Piner, K. R. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fisheries Research.

Cox, D.R., Snell, E.J., 1968. A general definition of residuals. J. R. Stat. Soc. Ser. B 30, 248–275.

Kell, L.T., Mosqueira, I., Grosjean, P., Fromentin, J.M., Garcia, D., Hillary, R., Jardim, E., Mardle, S., Pastoors, M.A., Poos, J.J. and Scott, F., 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES Journal of Marine Science, 64(4), pp.640-646.

Kell, L.T. and Mosqueira, I., 2017. Conditioning operating models on data and knowledge and rejecting and weighting of hypotheses. Collect. Vol. Sci. Pap. ICCAT, 73(9), pp.3000-3008.

Punt, A.E., A'mar, T., Bond, N.A., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A., Haltuch, M.A., Hollowed, A.B. and Szuwalski, C., 2013. Fisheries management under climate and environmental uncertainty: control rules and performance simulation. ICES Journal of Marine Science, 71(8), pp.2208-2220.

Rademeyer, R. A., Plagányi, É. E., & Butterworth, D. S. (2007). Tips and tricks in designing management procedures. ICES Journal of Marine Science, 64(4), 618-625.

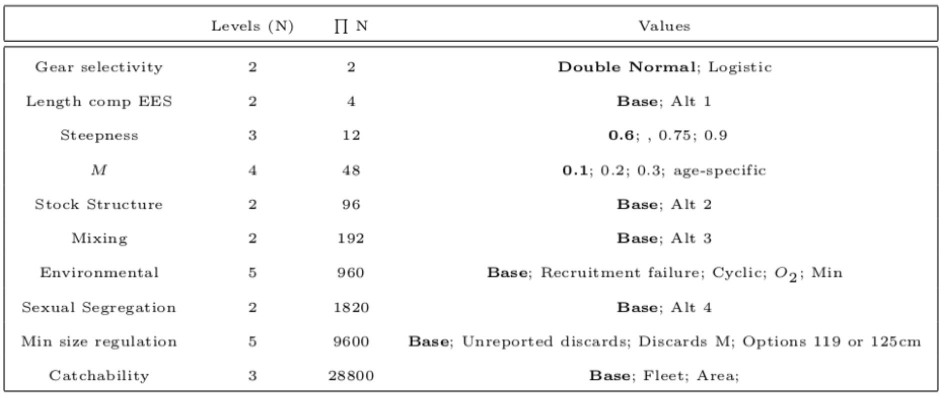
Schilling, M. F., 2012. The Surprising Predictability of Long Runs. Math. Mag. 85, 141-149

Sharma, R., Kitakado,T., Kell, L., Mosqueira, I., Kimoto, A., Scott, R., Davies, C.,, Kolody, D., Minte-Vera, C., Carruthers, T., Butterworth, D. and Miller, S. (submitted 2018). The current status of Operating Model Design in tRFMO's: Issues and lessons learned.

SEDAR 40, 2015. Atlantic Menhaden Stock Assessment Report. SEDAR, North Charleston, SC643.

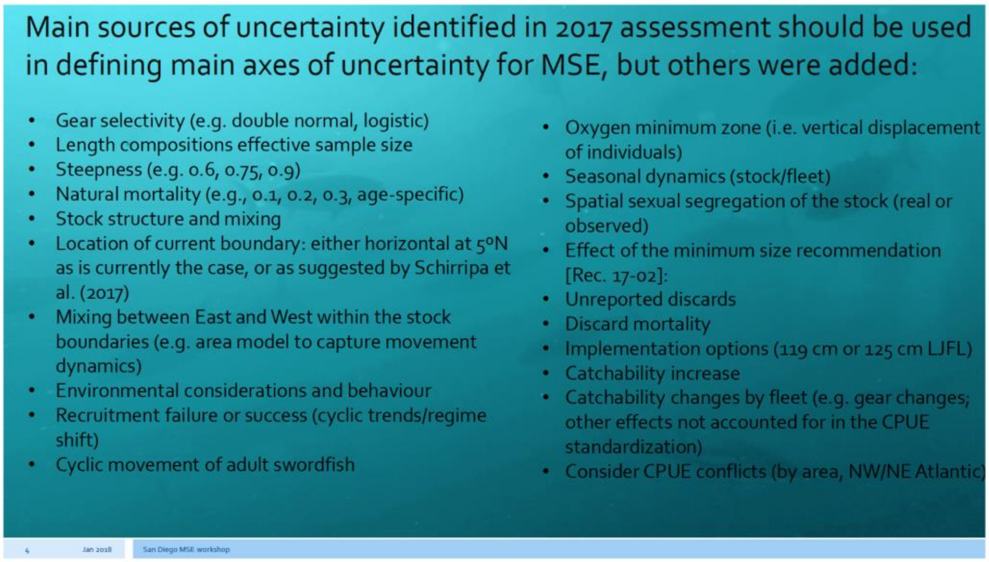
Wang, S.P., Chen, Y.R., Maunder, N.M., Nishida, T., 2009. Preliminary application of an age–structured assessment model to swordfish (Xiphias gladius) in the Indian Ocean. IOTC-WPB-2009-11, http://www.iotc.org/sites/default/files/documents/proceedings/2009/wpb/IOTC-2009-WPB-11.pdf.

Table 1: Operating Model Scenarios; Base Case values in bold.



	Levels (N)	∏ N	Values
Gear selectivity	2	2	**Double Normal**; Logistic
Length comp EES	2	4	**Base**; Alt 1
Steepness	3	12	**0.6**; , 0.75; 0.9
M	4	48	**0.1**; 0.2; 0.3; age-specific
Stock Structure	2	96	**Base**; Alt 2
Mixing	2	192	**Base**; Alt 3
Environmental	5	960	**Base**; Recruitment failure; Cyclic; O_2; Min
Sexual Segregation	2	1820	**Base**; Alt 4
Min size regulation	5	9600	**Base**; Unreported discards; Discards M; Options 119 or 125cm
Catchability	3	28800	**Base**; Fleet; Area;

Table 2. An expanded list of uncertainties



Main sources of uncertainty identified in 2017 assessment should be used in defining main axes of uncertainty for MSE, but others were added:

- Gear selectivity (e.g. double normal, logistic)
- Length compositions effective sample size
- Steepness (e.g. 0.6, 0.75, 0.9)
- Natural mortality (e.g., 0.1, 0.2, 0.3, age-specific)
- Stock structure and mixing
- Location of current boundary: either horizontal at 5ºN as is currently the case, or as suggested by Schirripa et al. (2017)
- Mixing between East and West within the stock boundaries (e.g. area model to capture movement dynamics)
- Environmental considerations and behaviour
- Recruitment failure or success (cyclic trends/regime shift)
- Cyclic movement of adult swordfish
- Oxygen minimum zone (i.e. vertical displacement of individuals)
- Seasonal dynamics (stock/fleet)
- Spatial sexual segregation of the stock (real or observed)
- Effect of the minimum size recommendation [Rec. 17-02]:
- Unreported discards
- Discard mortality
- Implementation options (119 cm or 125 cm LJFL)
- Catchability increase
- Catchability changes by fleet (e.g. gear changes; other effects not accounted for in the CPUE standardization)
- Consider CPUE conflicts (by area, NW/NE Atlantic)

4 Jan 2018 San Diego MSE workshop

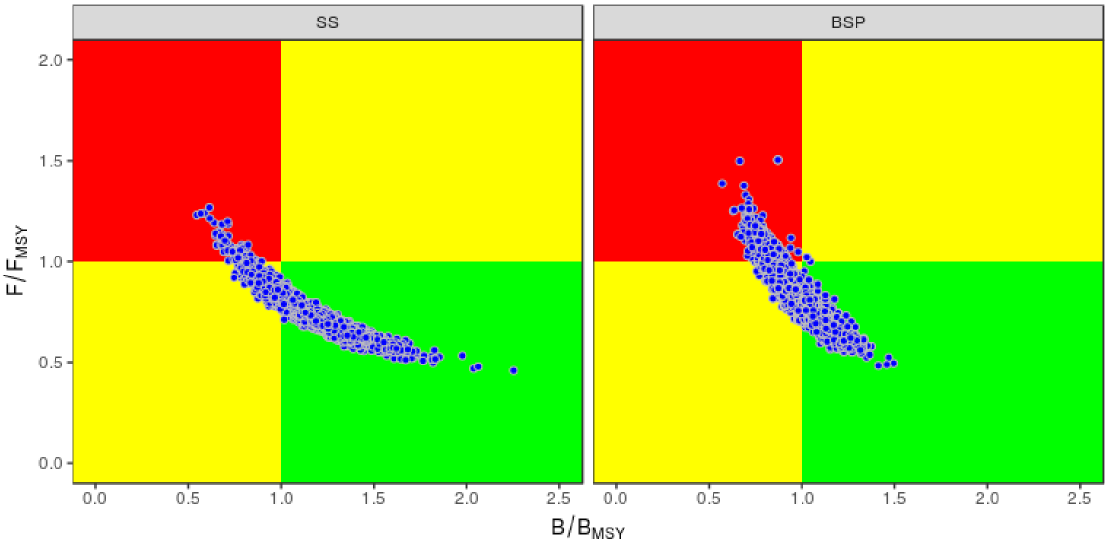


Figure 1. Kobe phase plot for North Atlantic swordfish from 2017 stock assessment by assessment method.

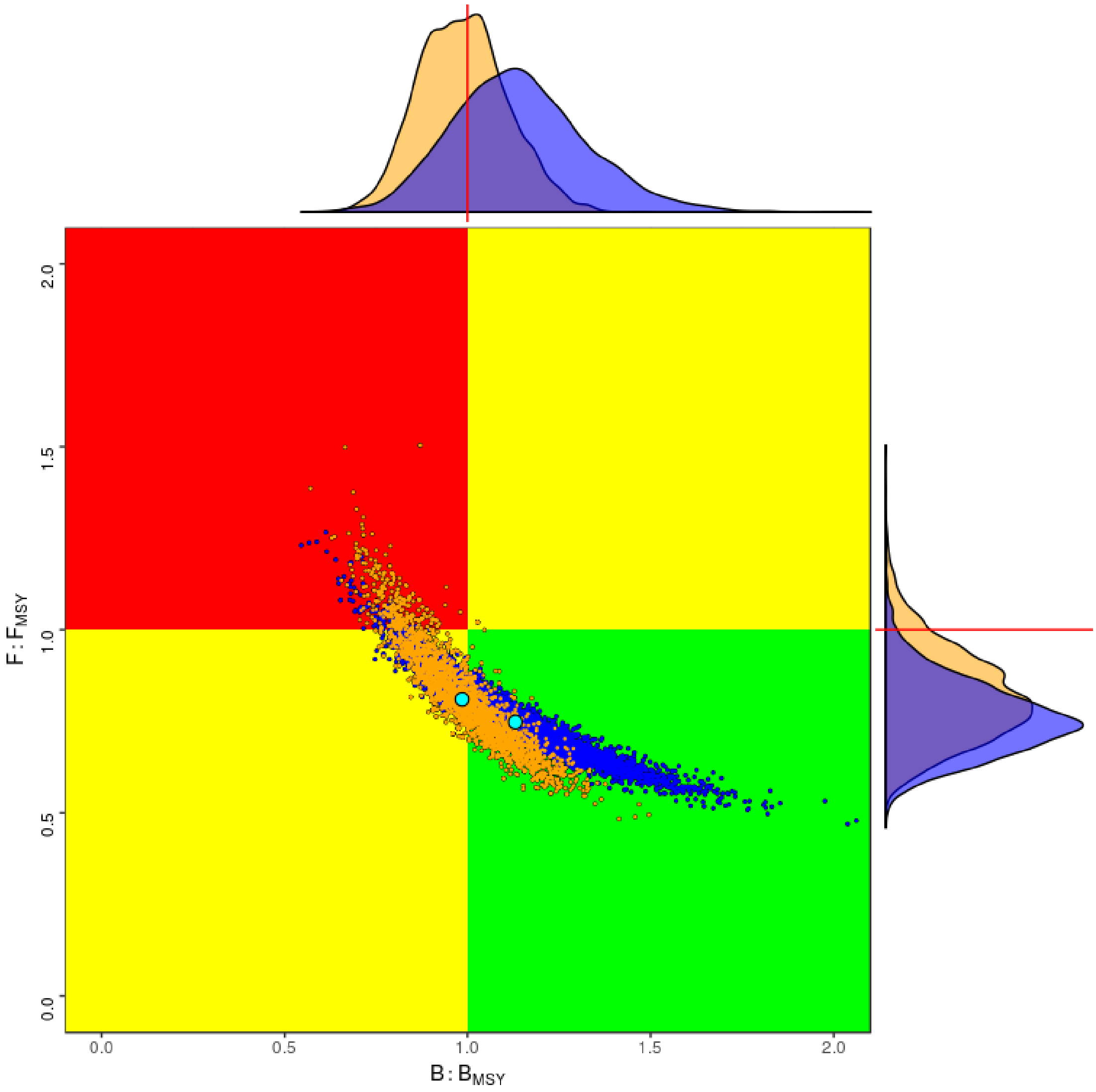


Figure 2. Kobe phase plot for North Atlantic swordfish from 2017 stock assessment by assessment method combined with marginal distributions, cyan points show the medians of the two distributions.

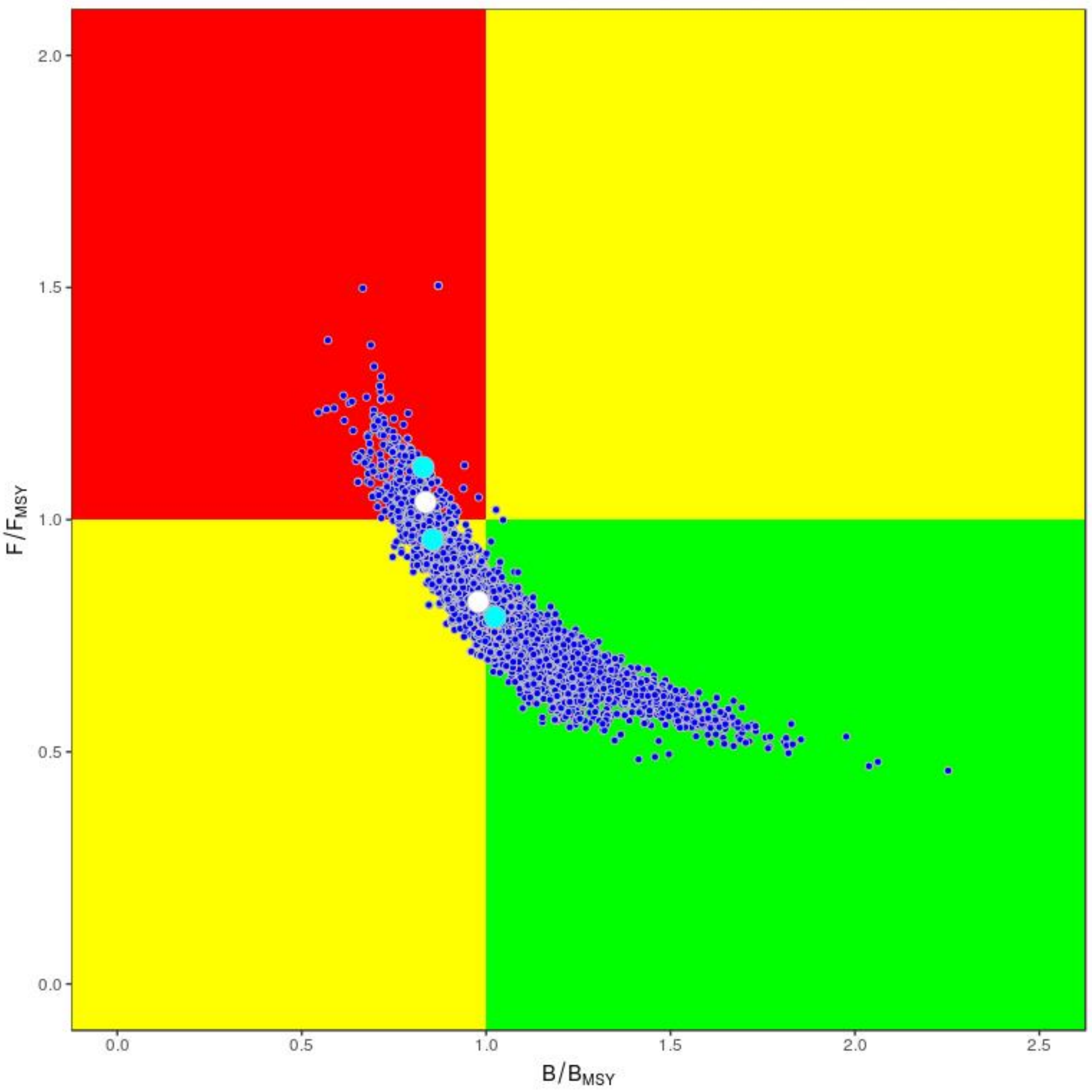


Figure 3. Kobe phase plot North.

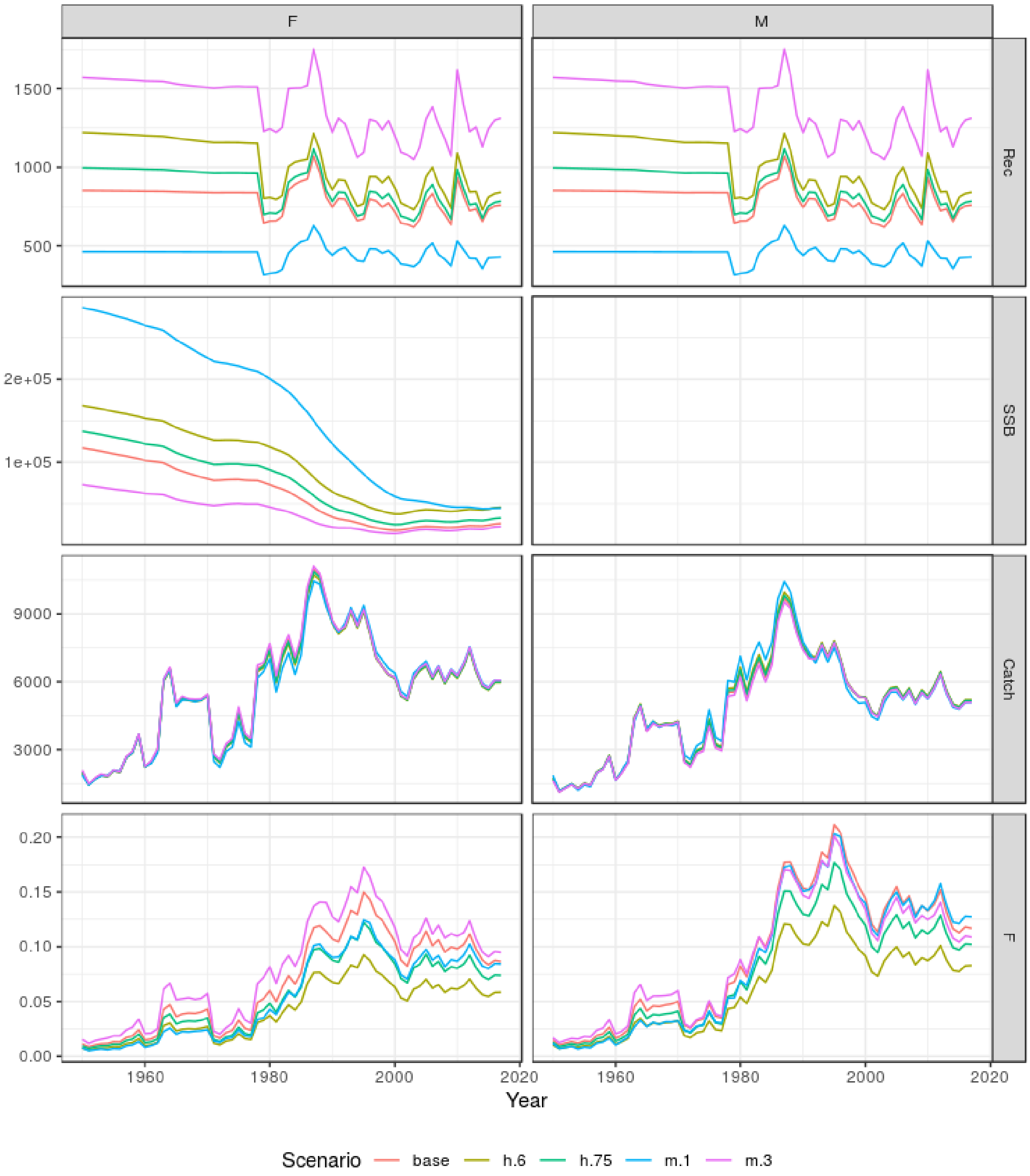


Figure 4. Time series of recruitment, SSB, catch and fishing mortality.

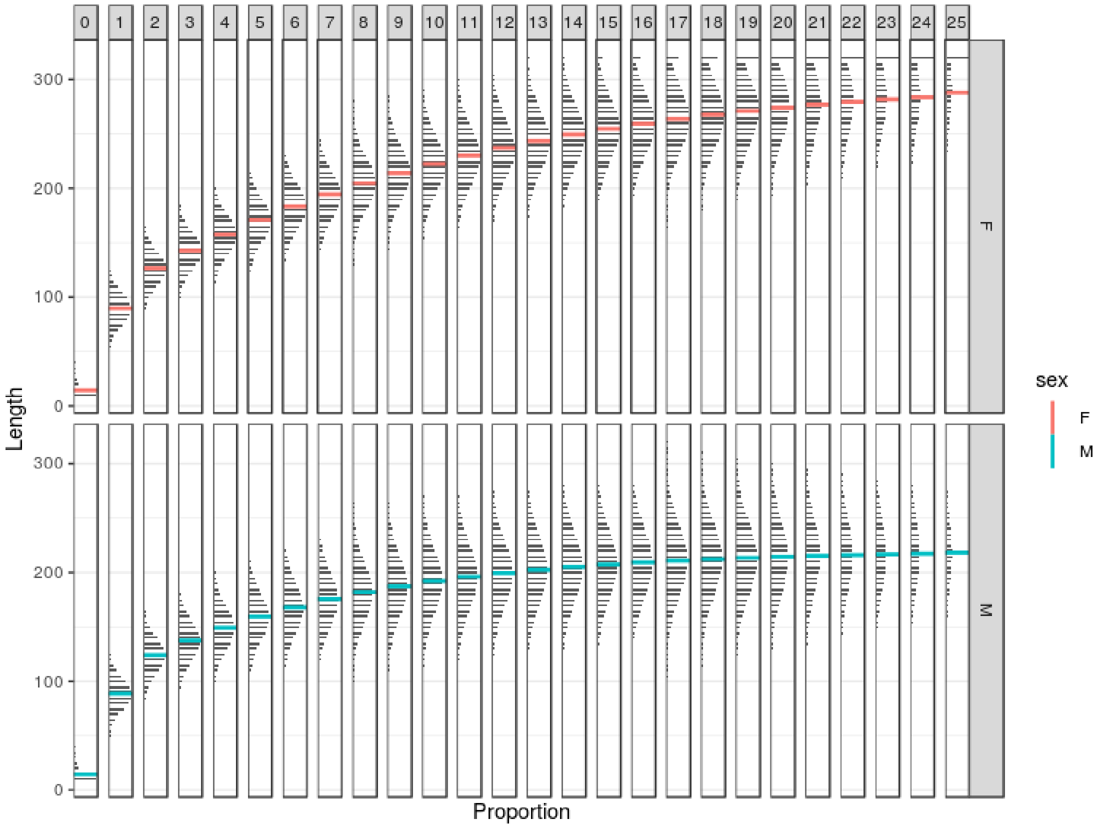


Figure 5. Length distributions by age from the age length key generated by SS3 for scenario **base**.

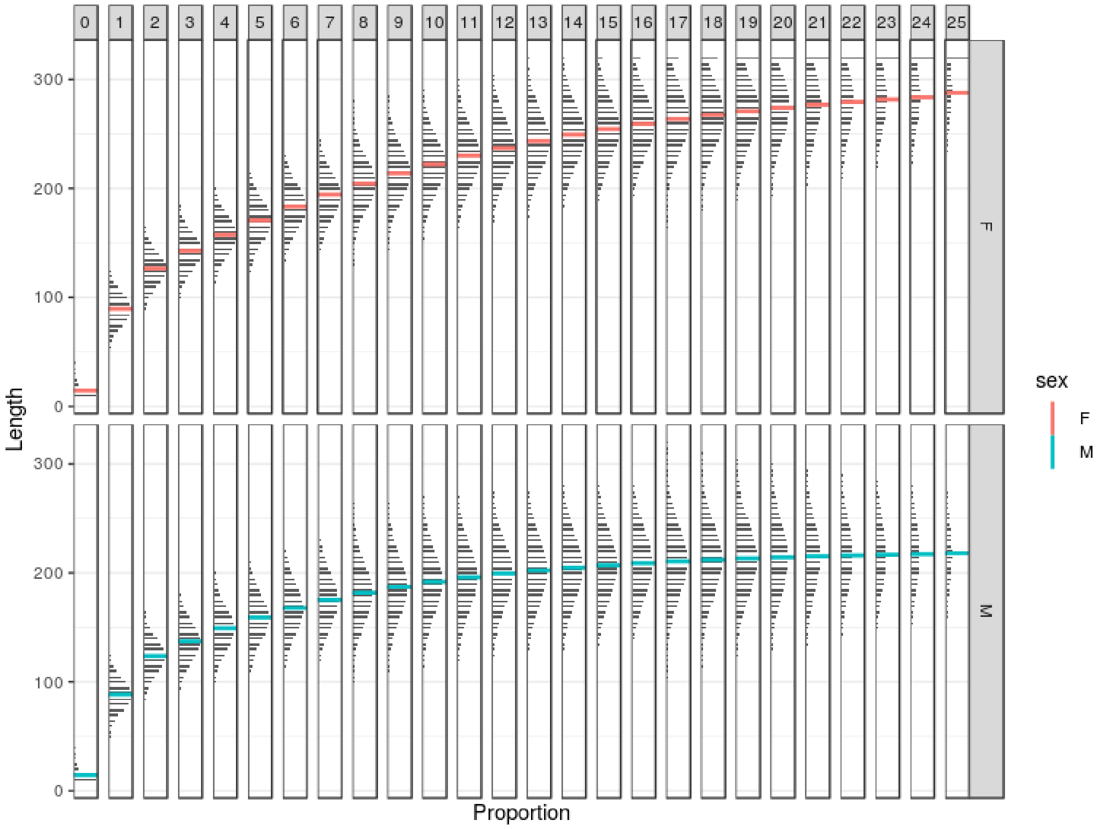


Figure 6. Length distributions by age from the age length key generated by SS3 for scenario **h.6**.

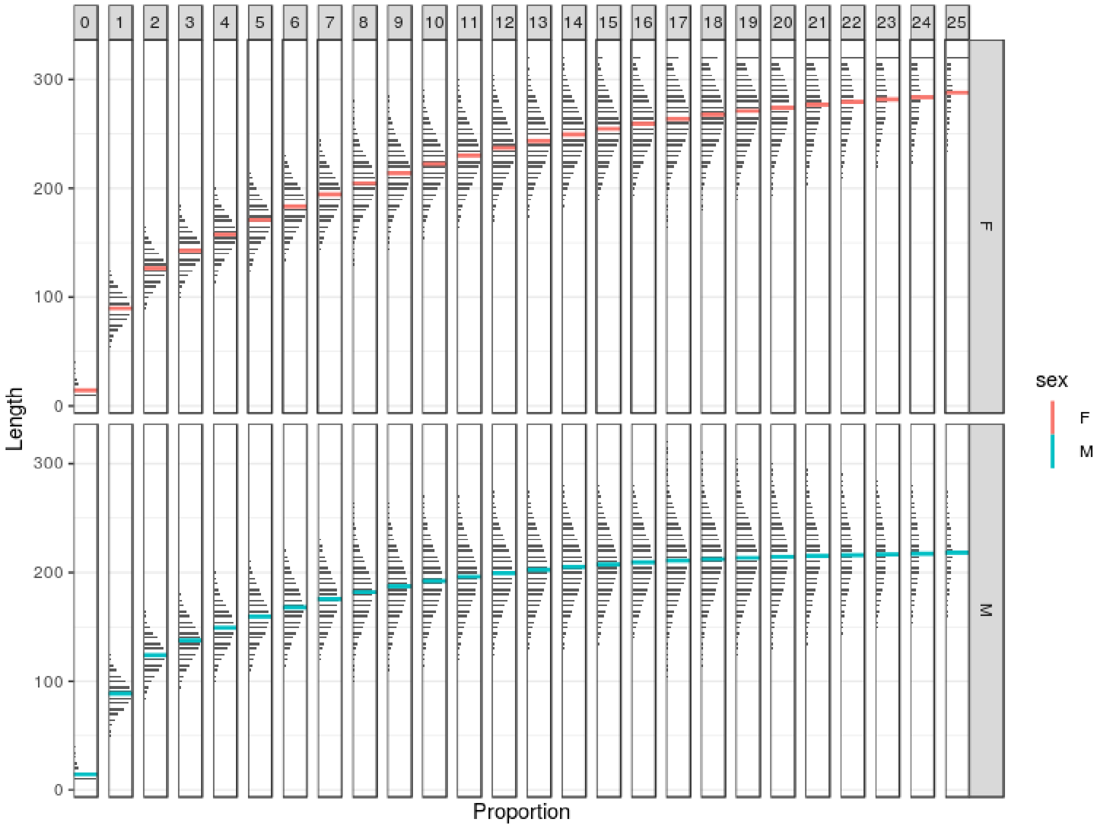


Figure 7. Length distributions by age from the age length key generated by SS3 for scenario **h.75**.

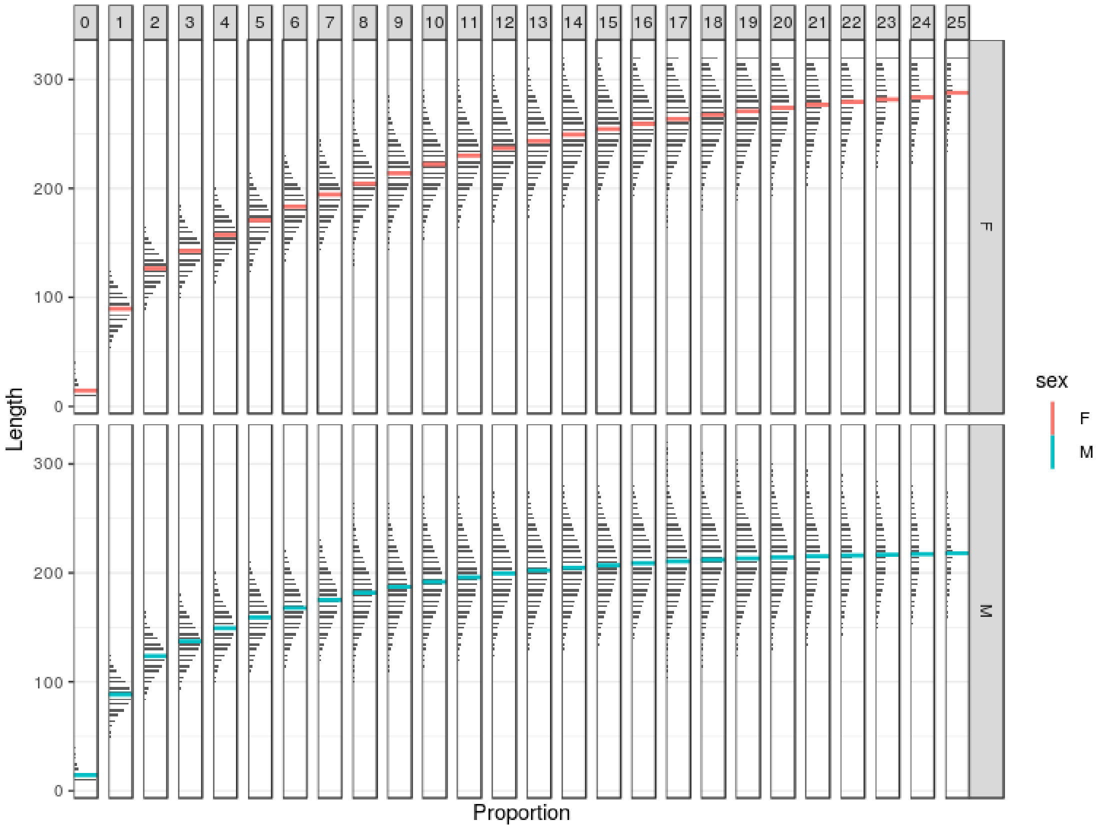


Figure 8. Length distributions by age from the age length key generated by SS3 for scenario **m.1**.

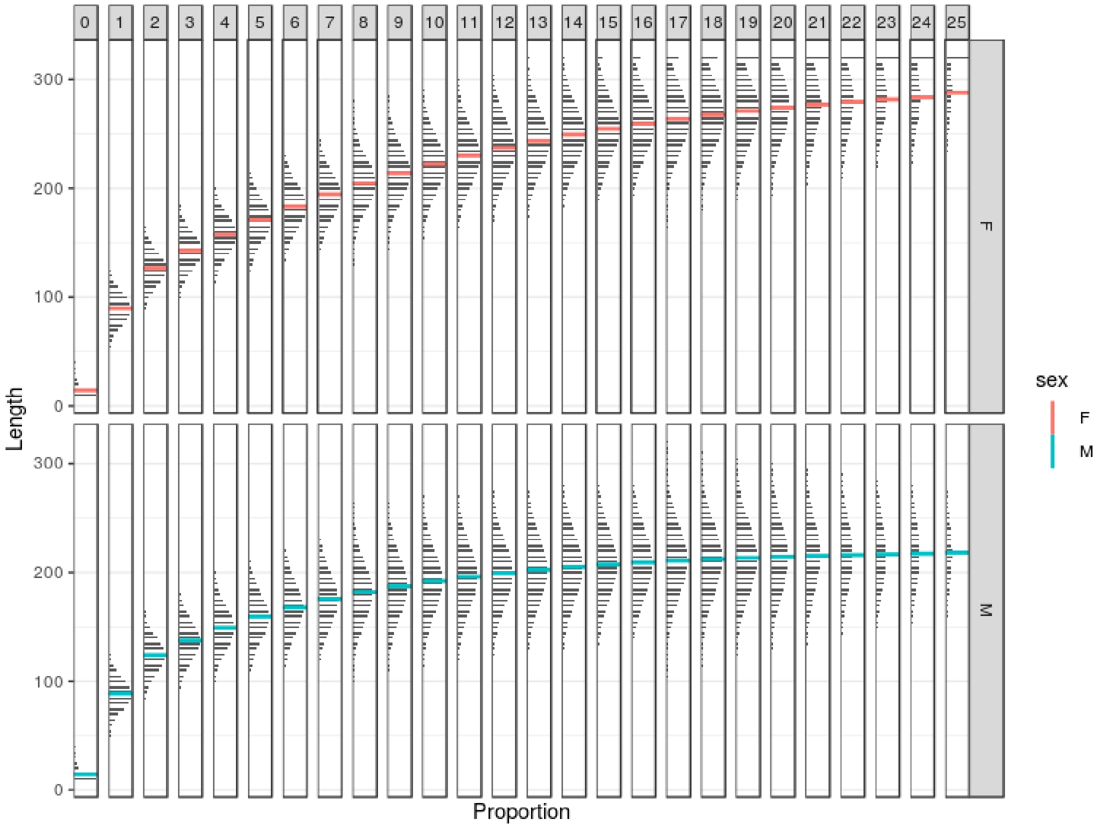


Figure 9. Length distributions by age from the age length key generated by SS3 for scenario **m.3**.

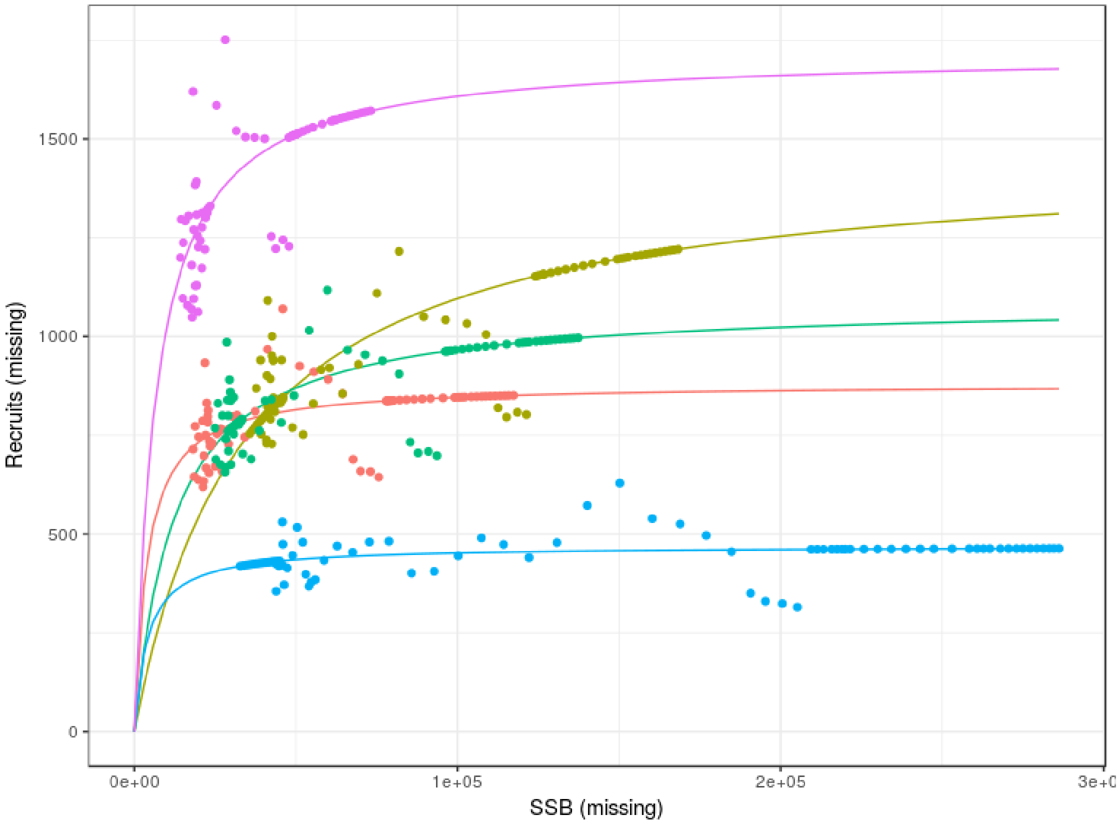


Figure 10. Stock recruitment relationship as fitted by SS.

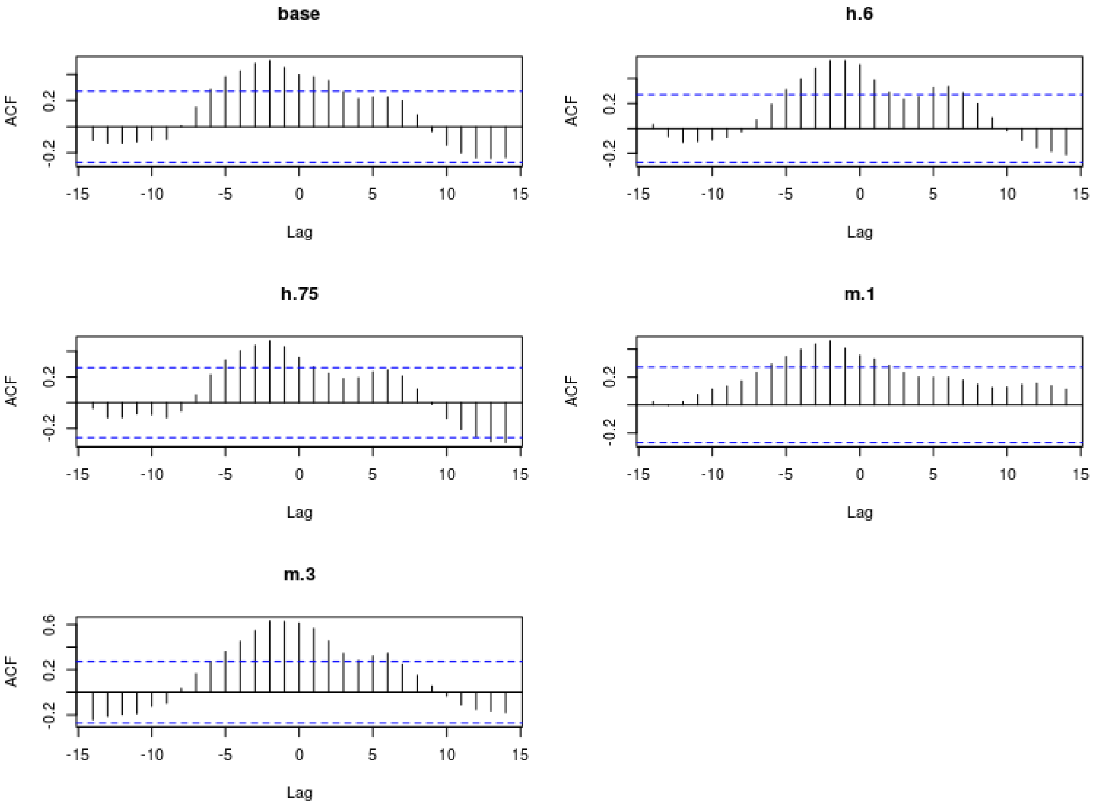


Figure 11. Cross Spearman correlations between SSB and recruitment, a positive lag indicates that a stock-recruit relationship and a negative lag a recruit-stock relationship.

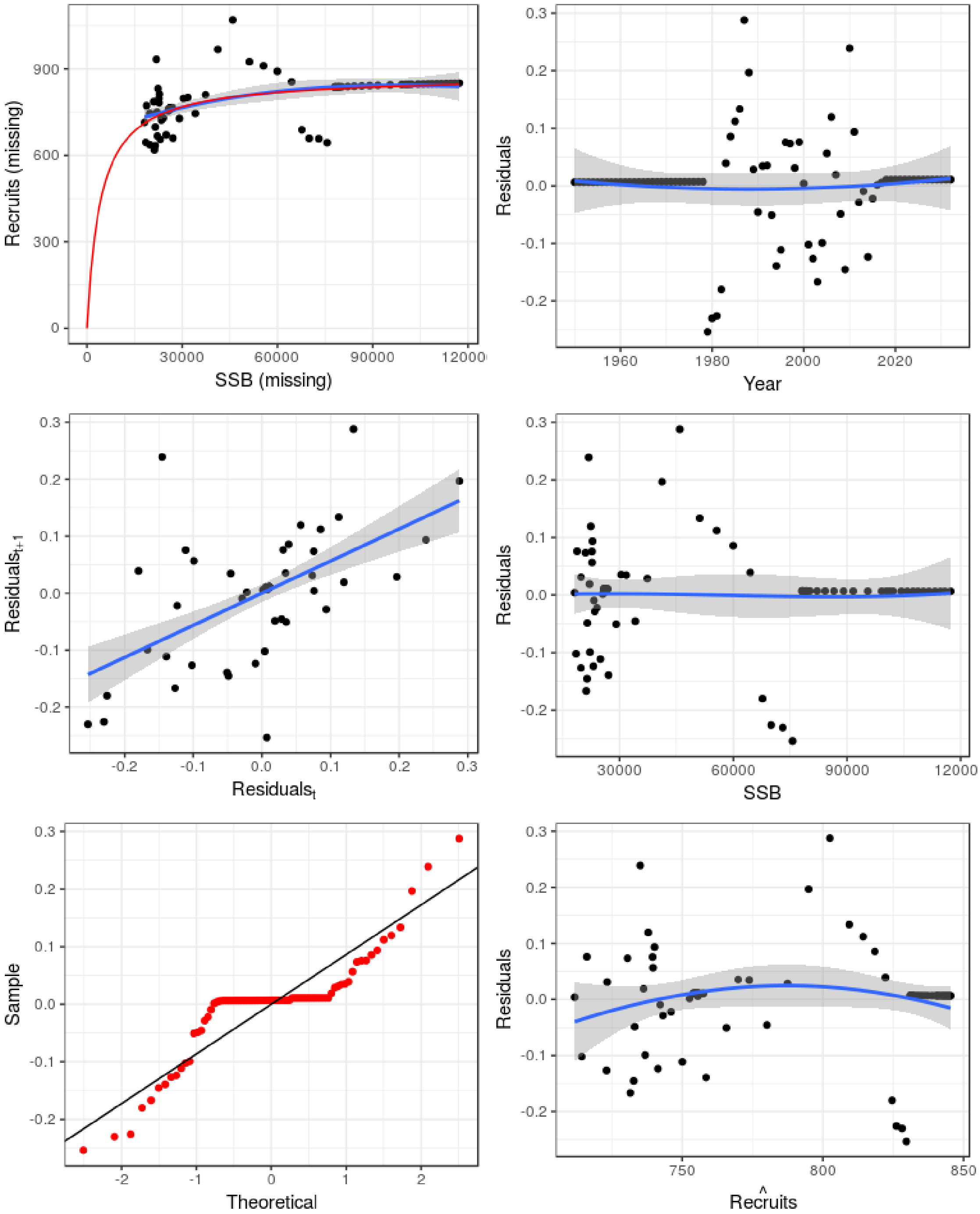


Figure 12. Fit to stock recruitment relationship for scenario **base** with goodness of fit diagnostics.

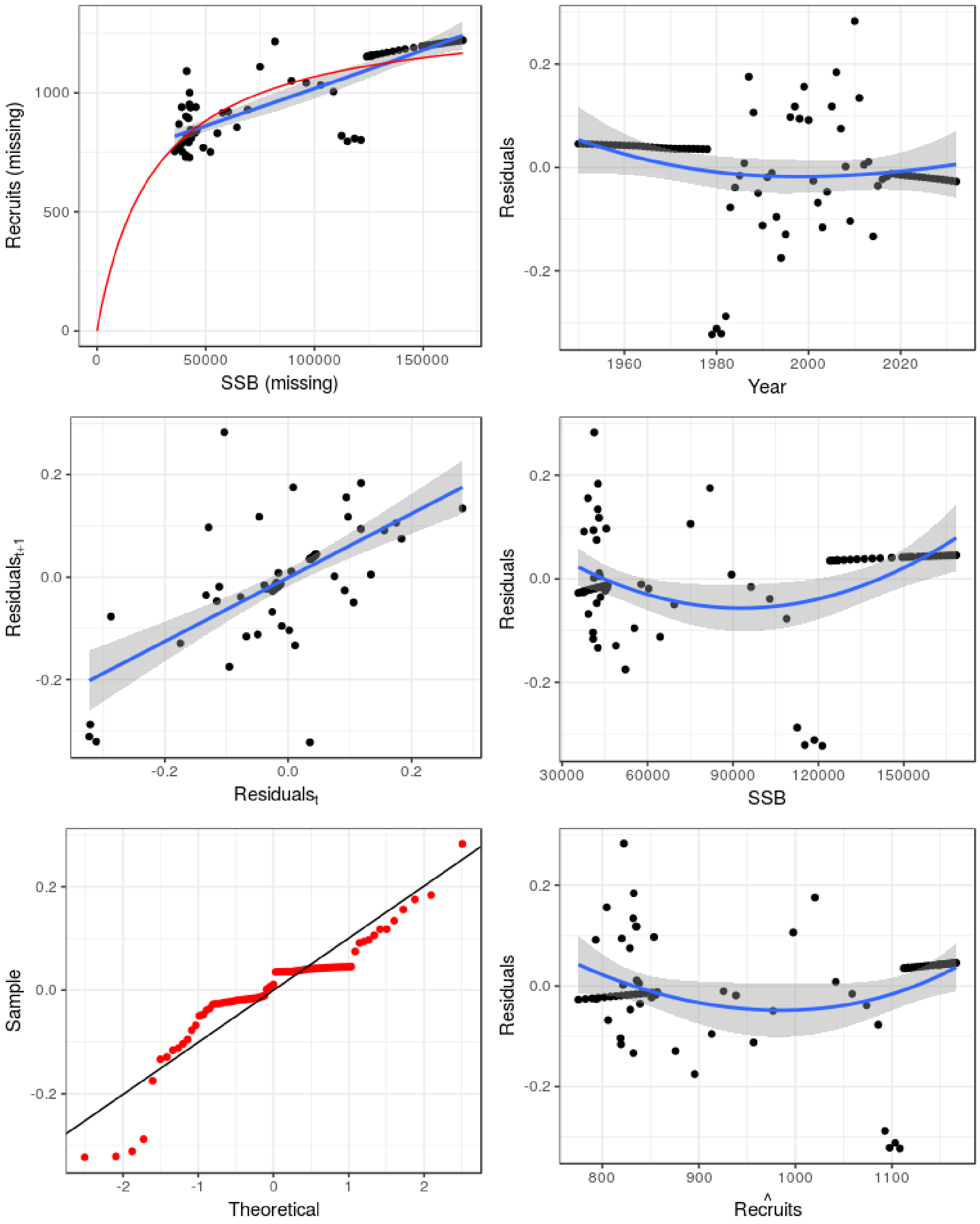


Figure 13. Fit to stock recruitment relationship for scenario **h.6.**, with goodness of fit diagnostics.

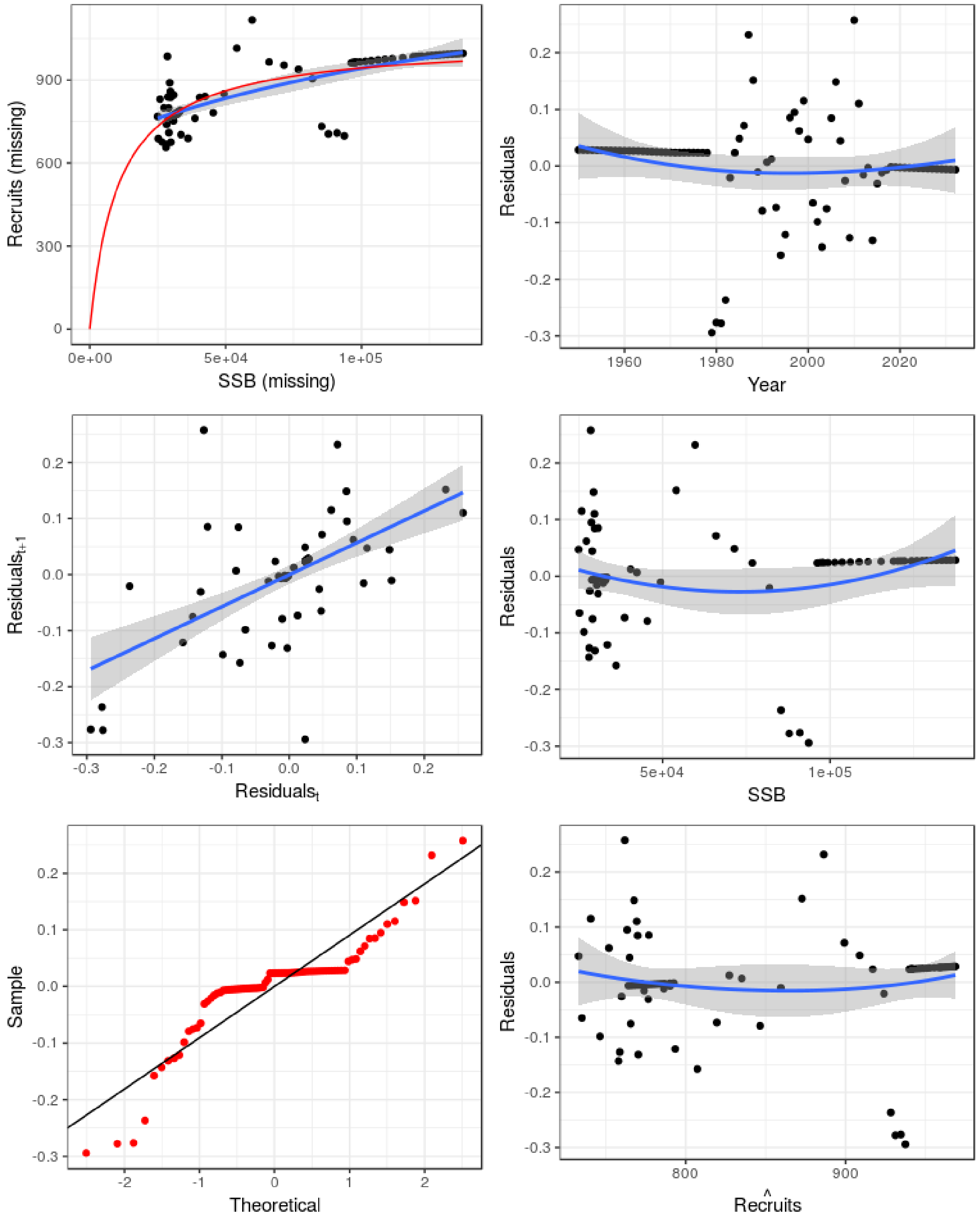


Figure 14. Fit to stock recruitment relationship for scenario **h.75.**, with goodness of fit diagnostics.

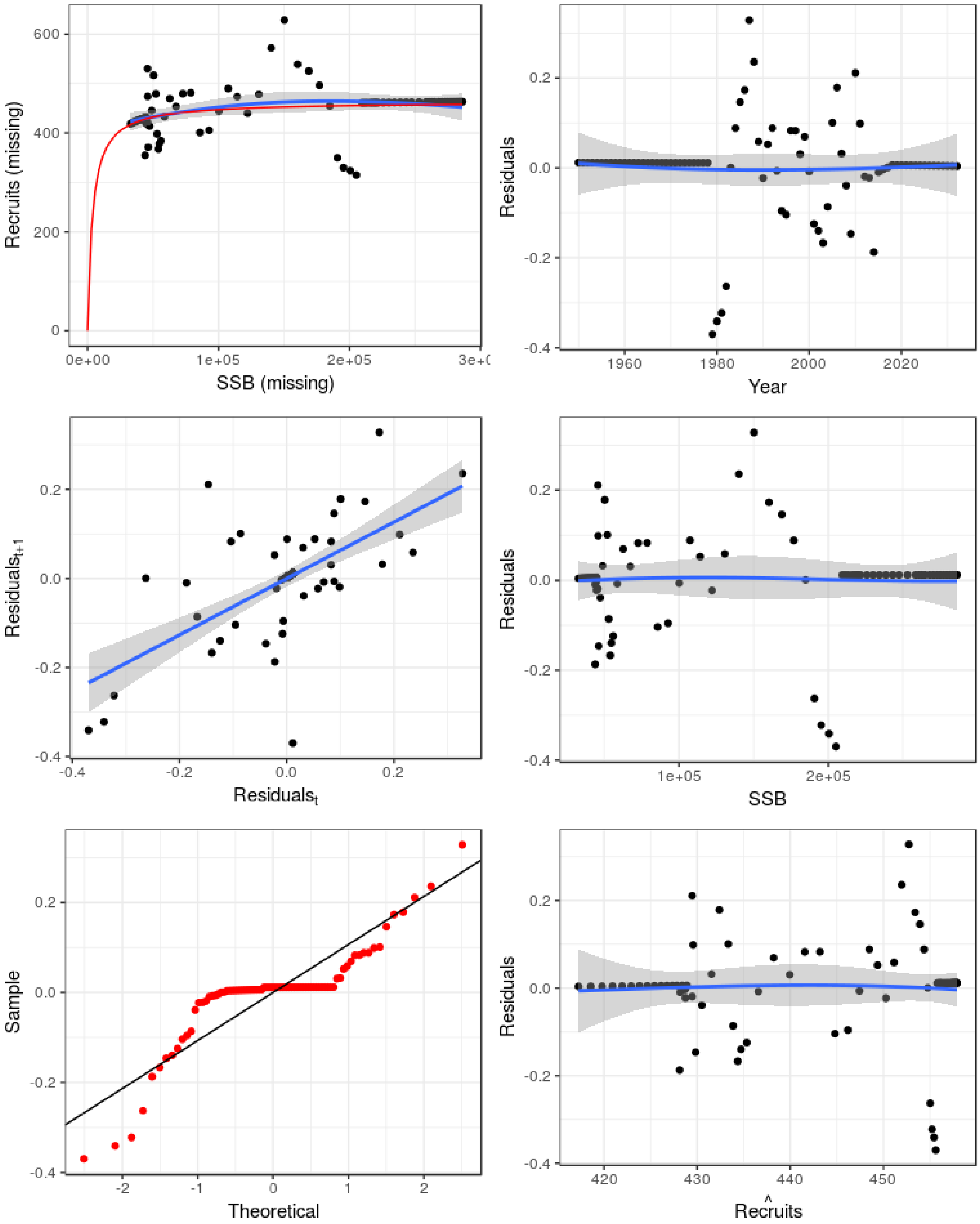


Figure 15. Fit to stock recruitment relationship for scenario **m.1**., with goodness of fit diagnostics.

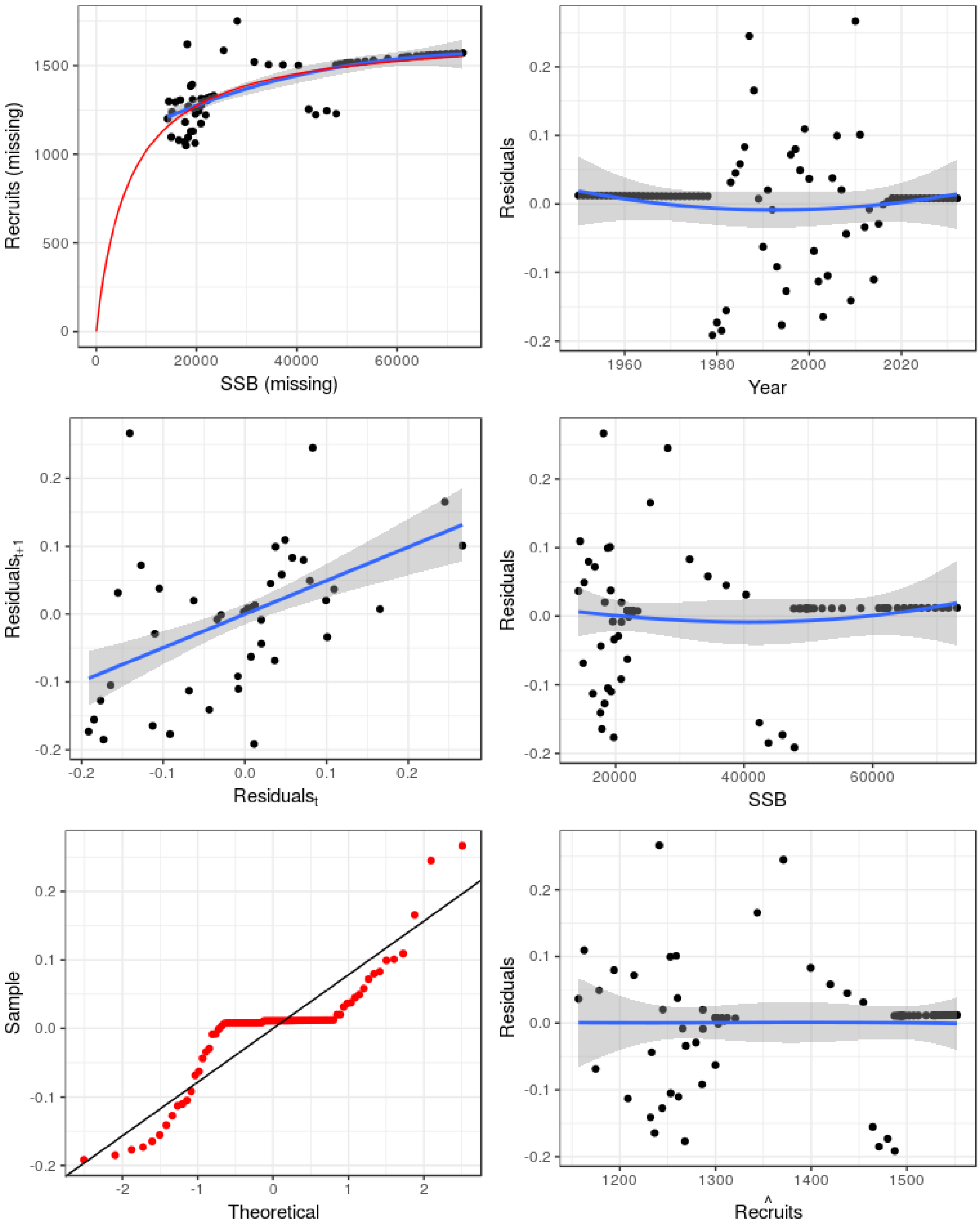


Figure 16. Fit to stock recruitment relationship for scenario **m.3**., with goodness of fit diagnostics.

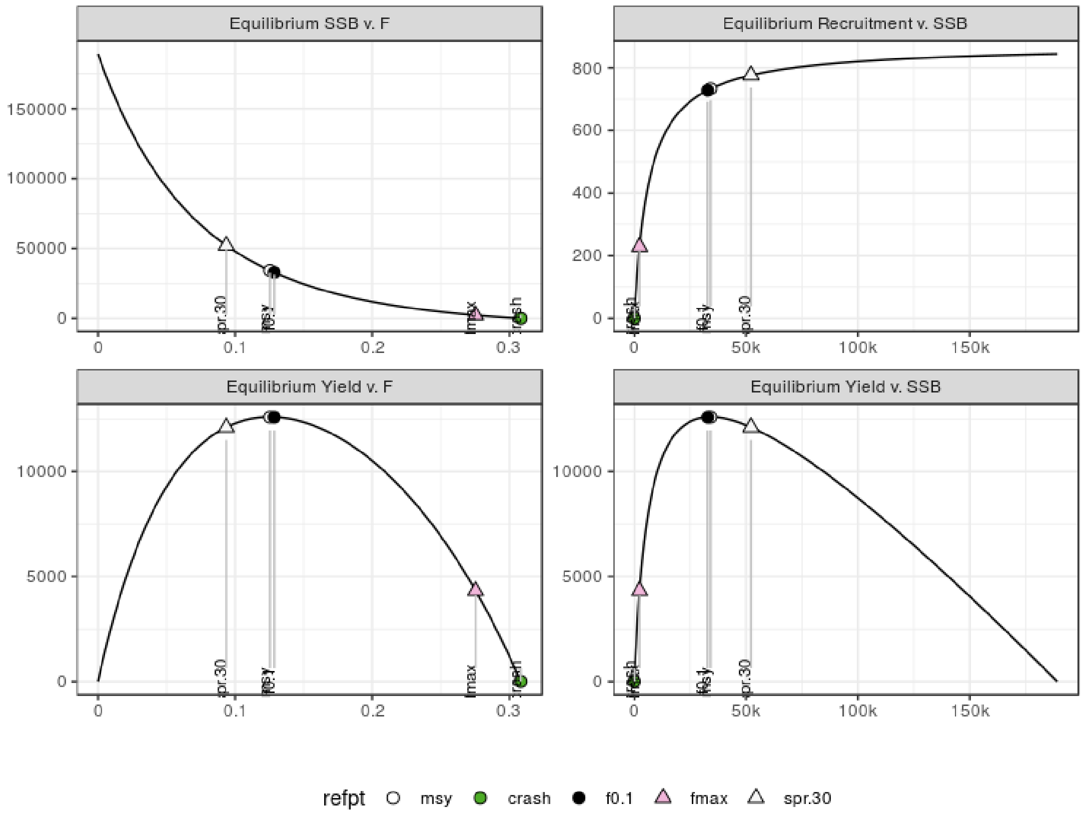


Figure 17. Equilibrium curves with reference points for scenario **base**.

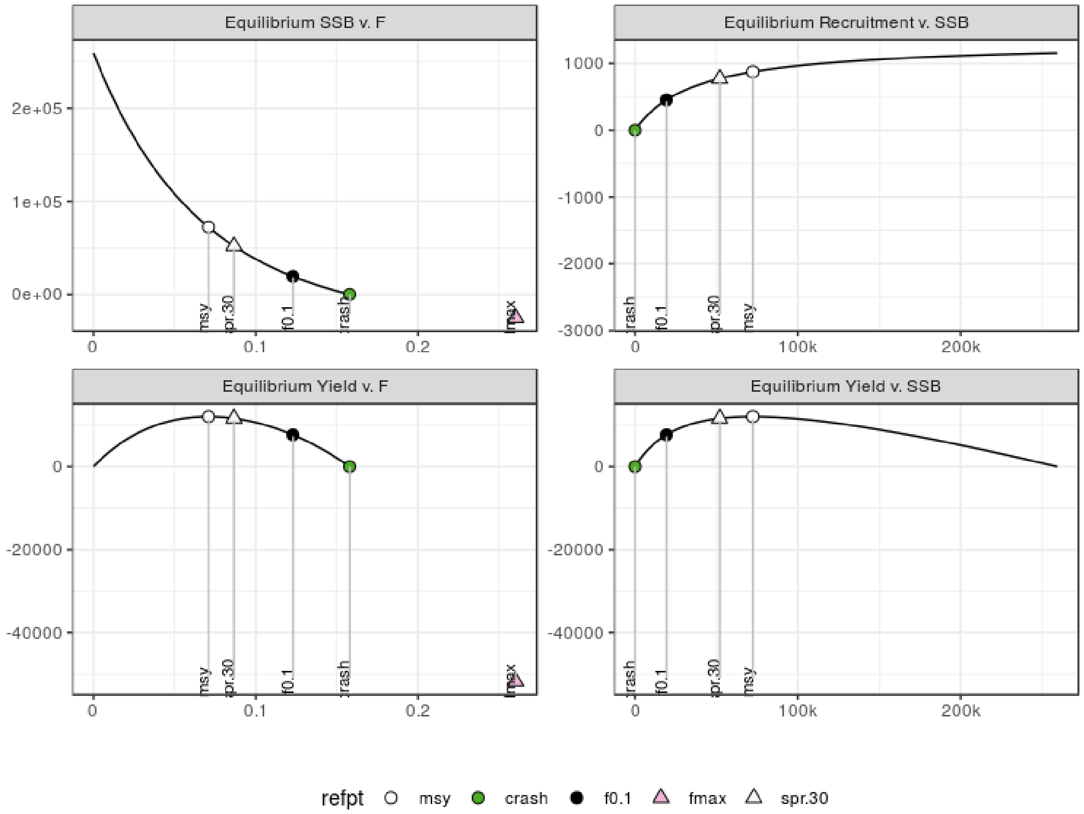


Figure 18. Equilibrium curves with reference points for scenario **h.6**.

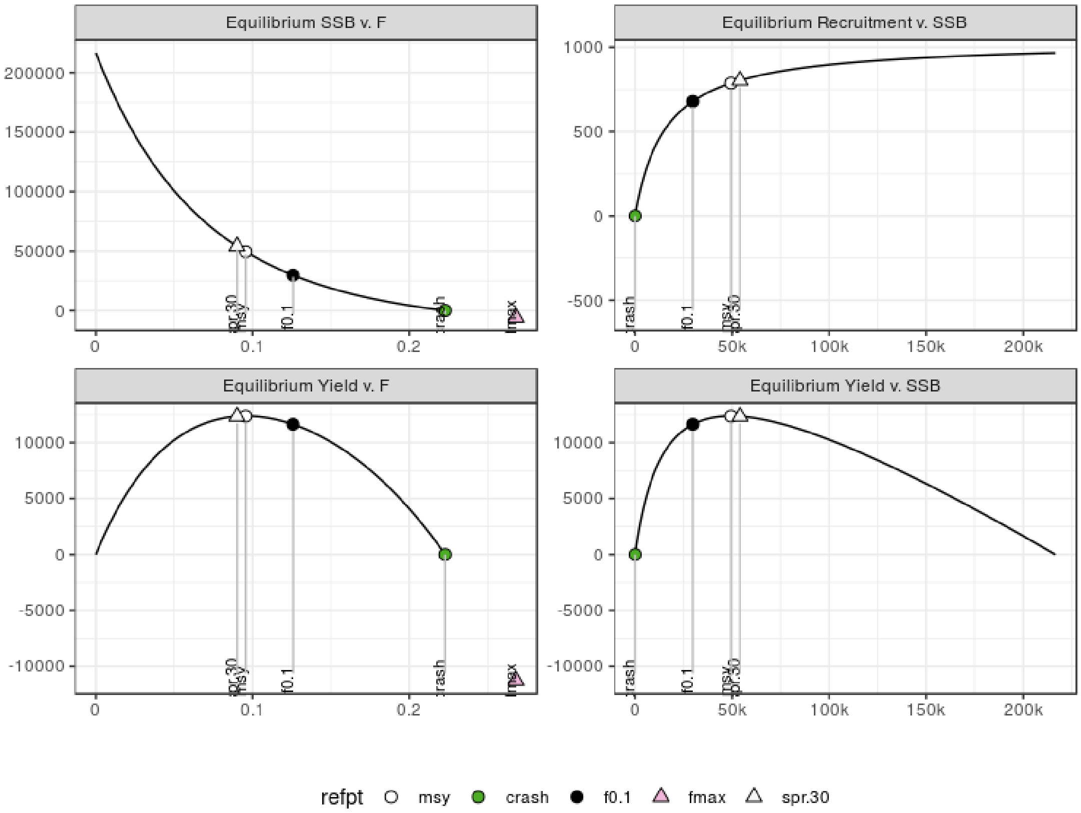


Figure 19. Equilibrium curves with reference points for scenario **h.75**.

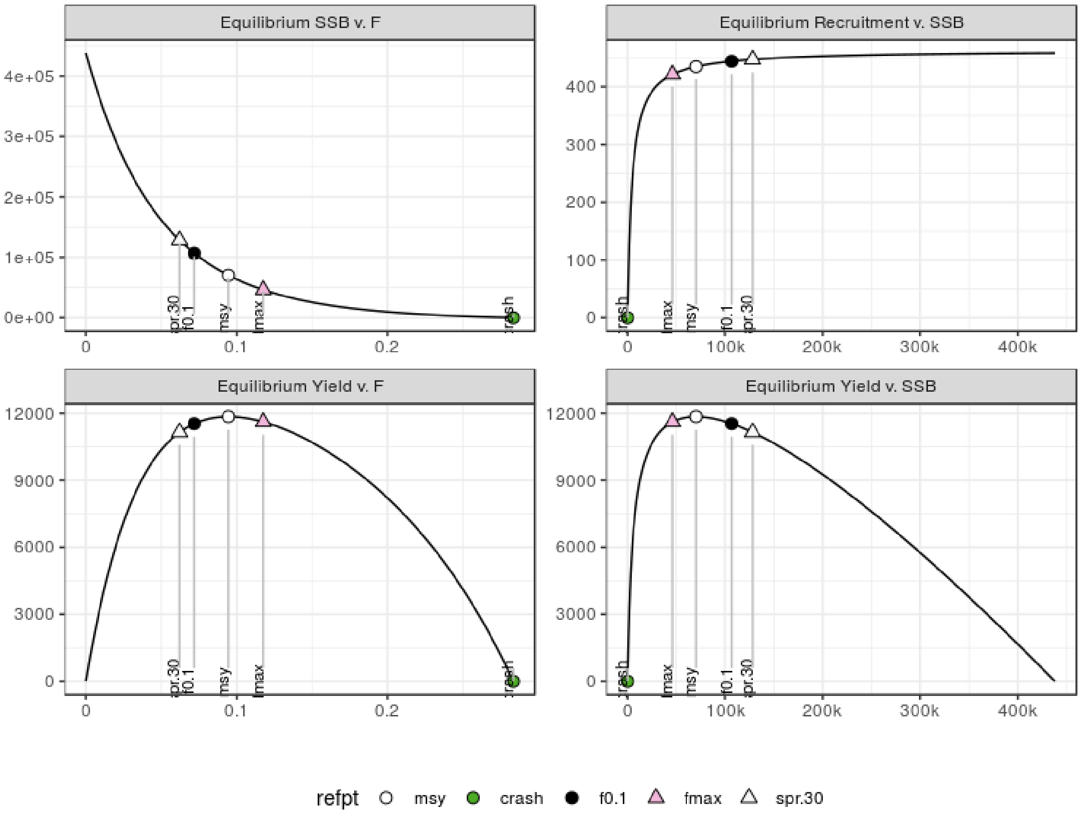


Figure 20. Equilibrium curves with reference points for scenario **m.1**.

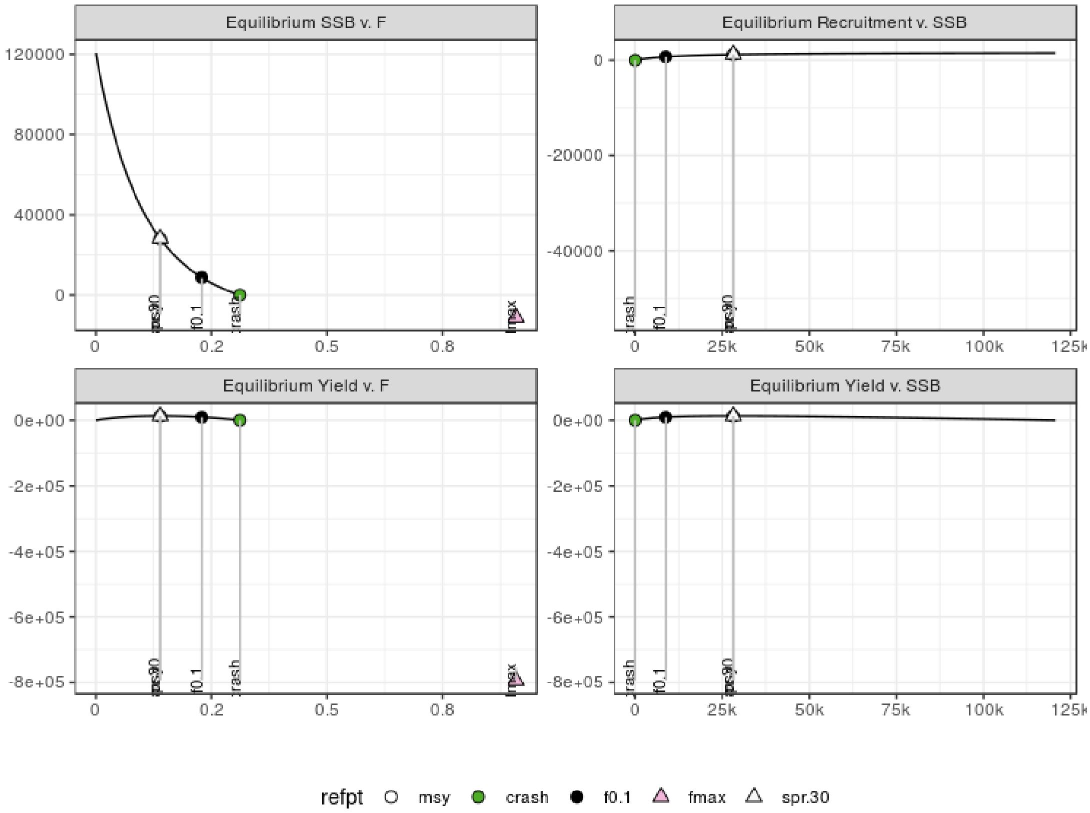


Figure 21. Equilibrium curves with reference points for scenario **m.3**.

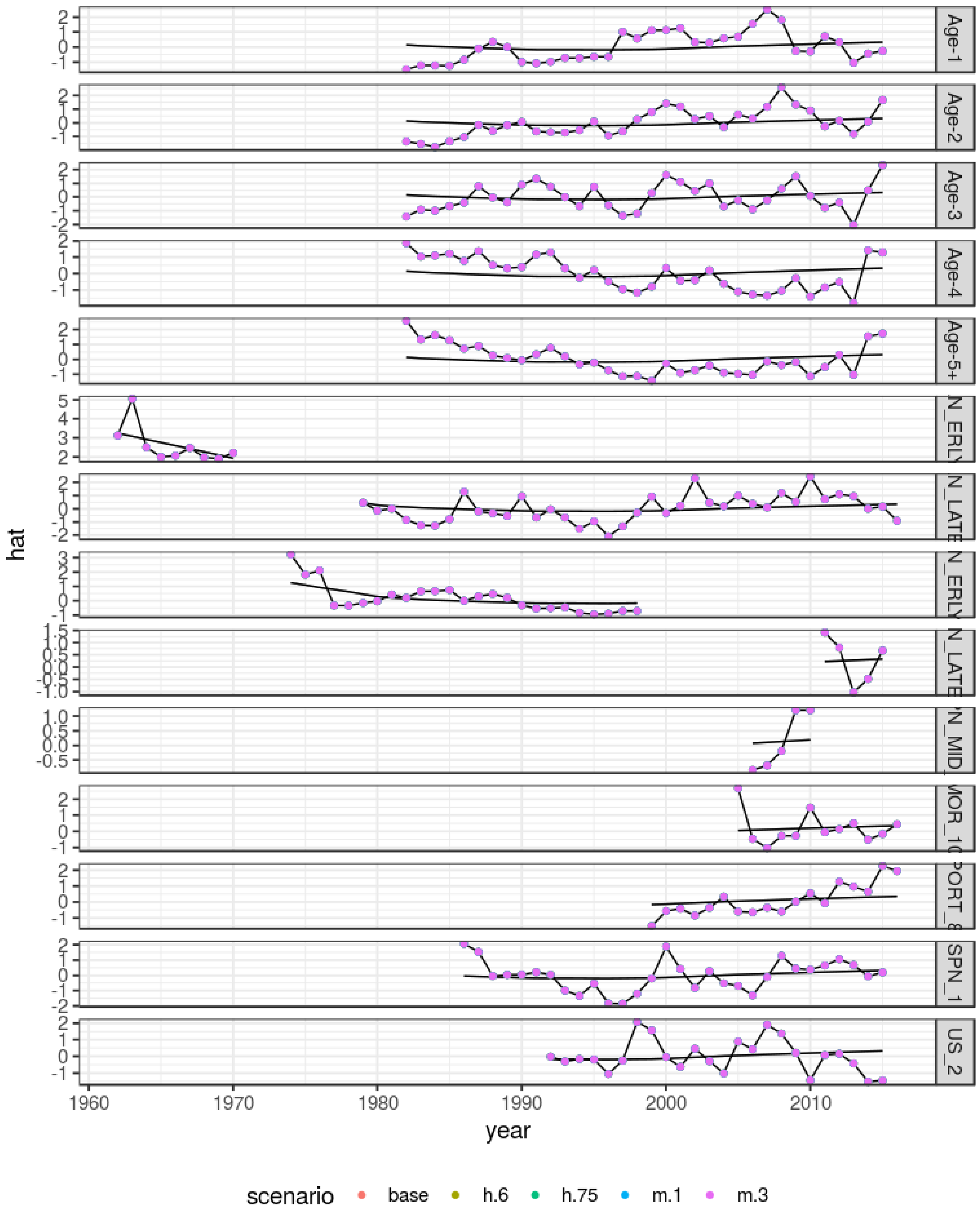


Figure 22. Time series of CPUE indices, continuous black line is a lowess smother showing the average trend by area (i.e. fitted to year for each area with series as a factor)

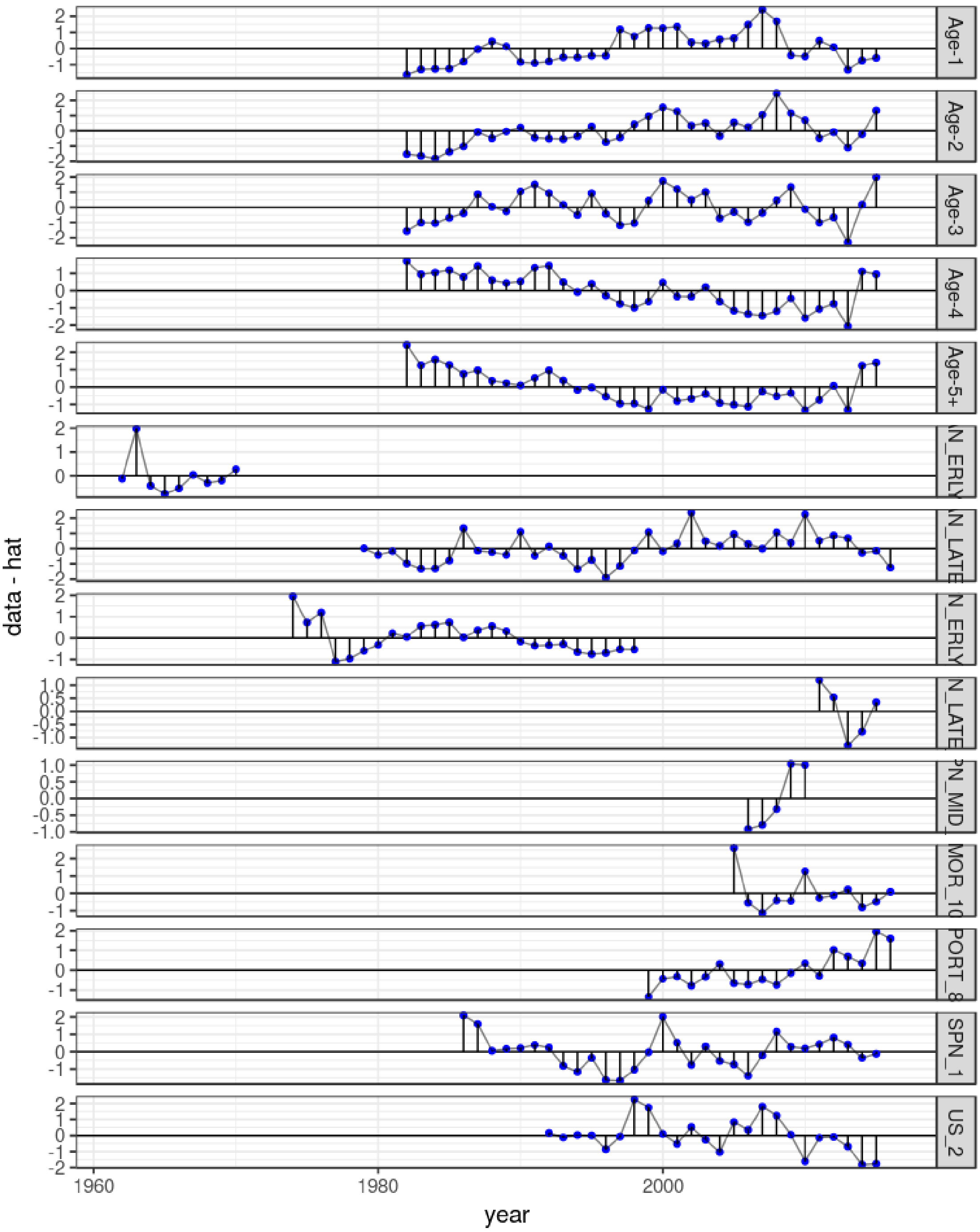


Figure 23. Time series of residuals from the lowess fit.

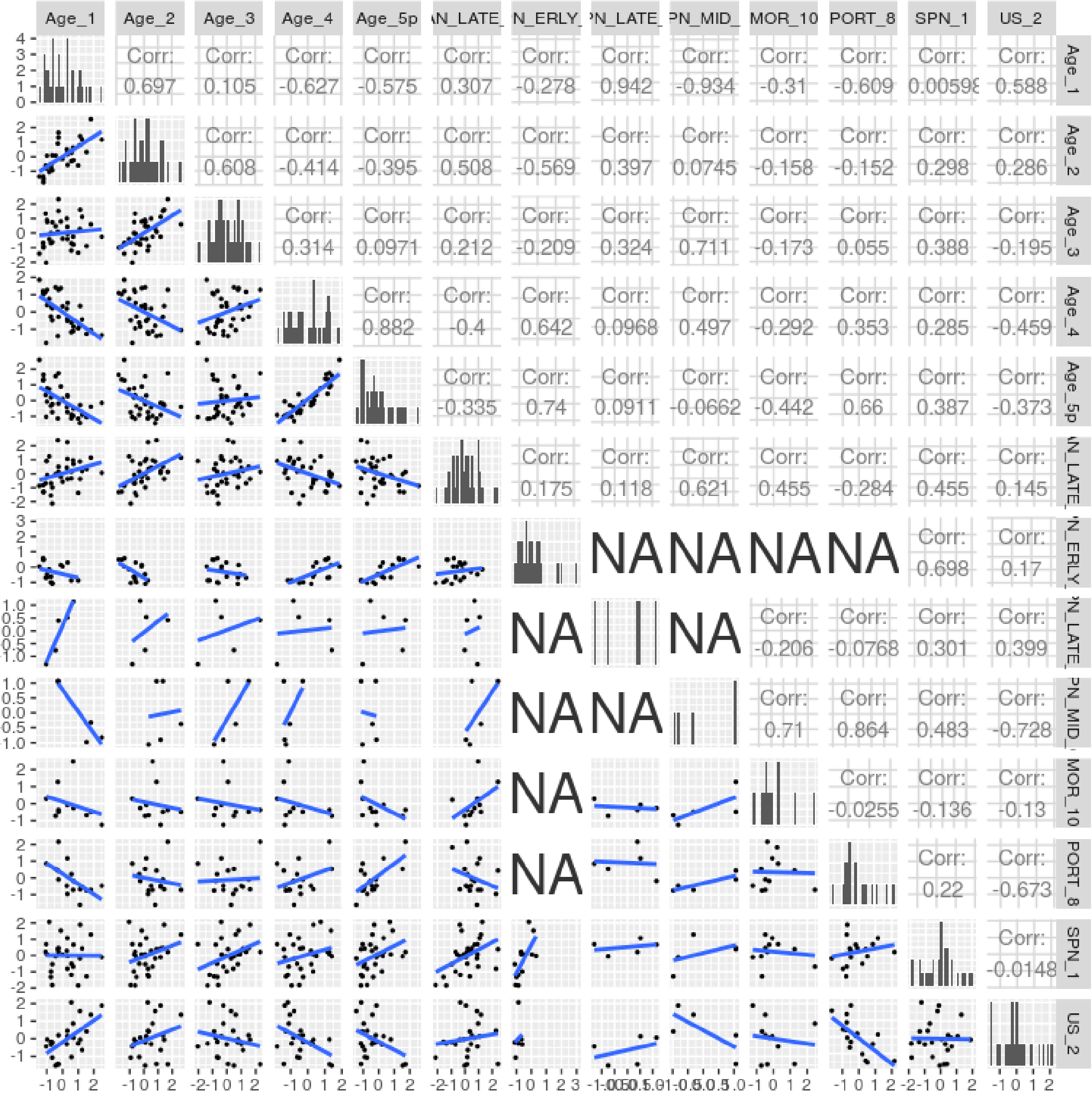


Figure 24. Pairwise scatter plots to look at correlations between Indices.

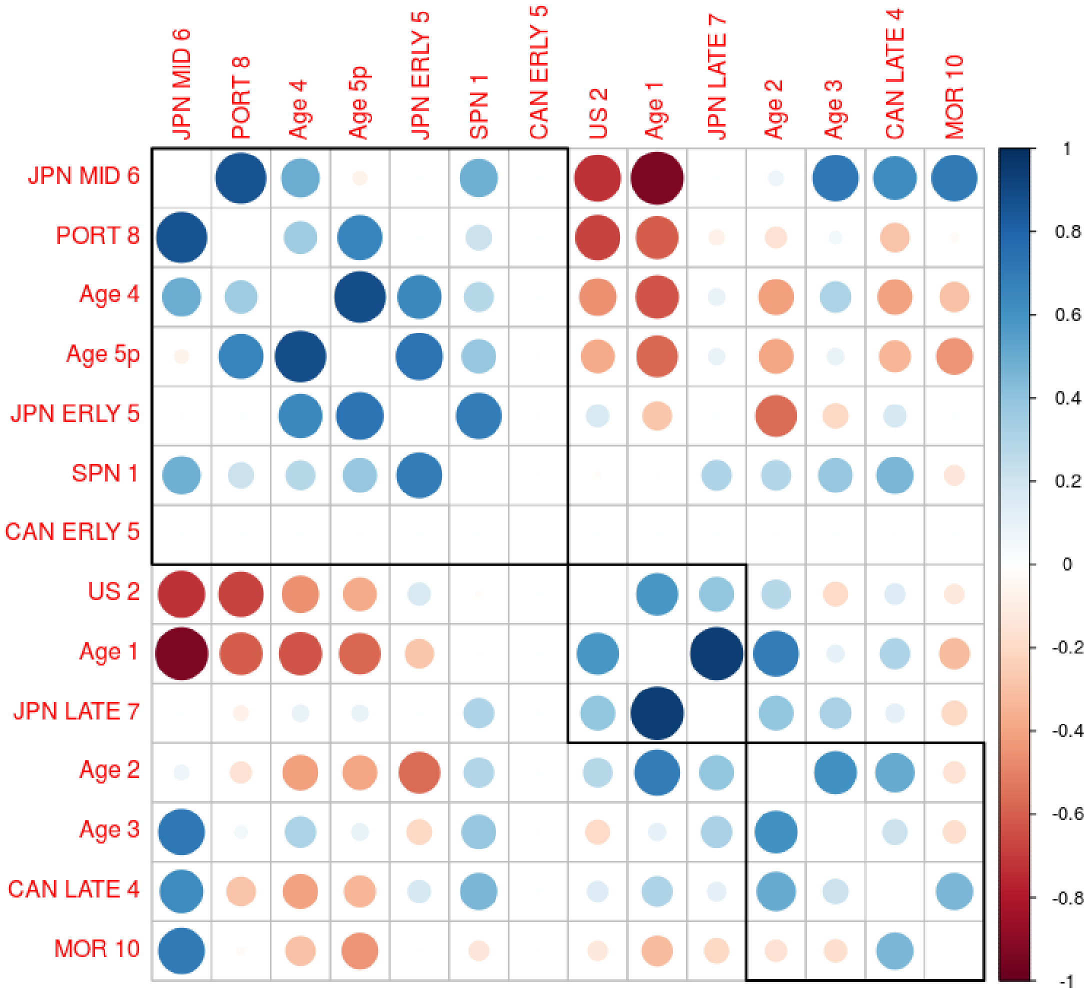


Figure 25. Plot of the correlation matrix for the CPUE indices, blue indicate a positive correlation and red negative. The order of the indices and the rectangular boxes are chosen based on a hierarchical cluster analysis using a set of dissimilarities for the indices being clustered.

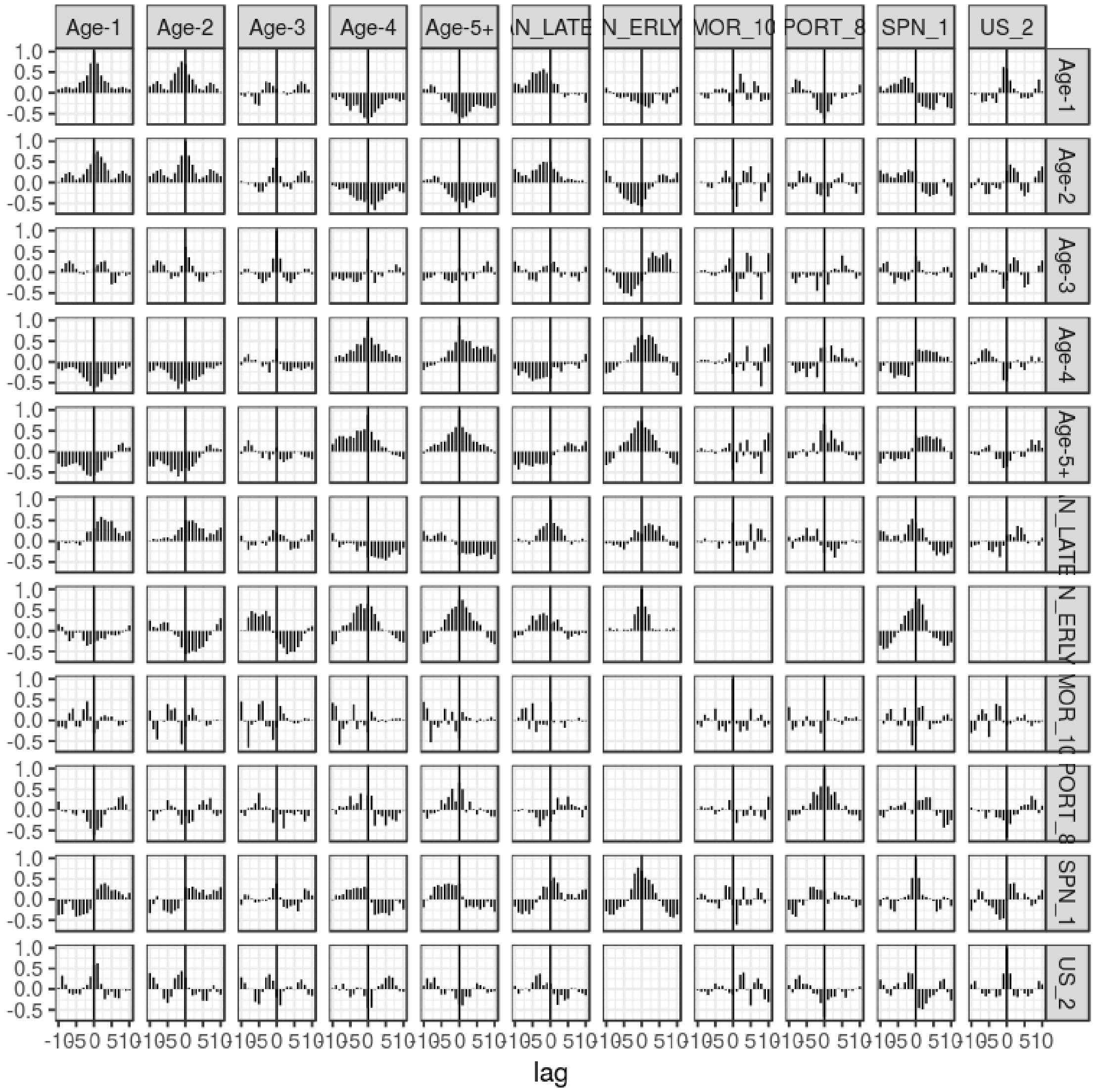


Figure 26. Cross correlations between indices, to identify potential lags due to year-class effects.

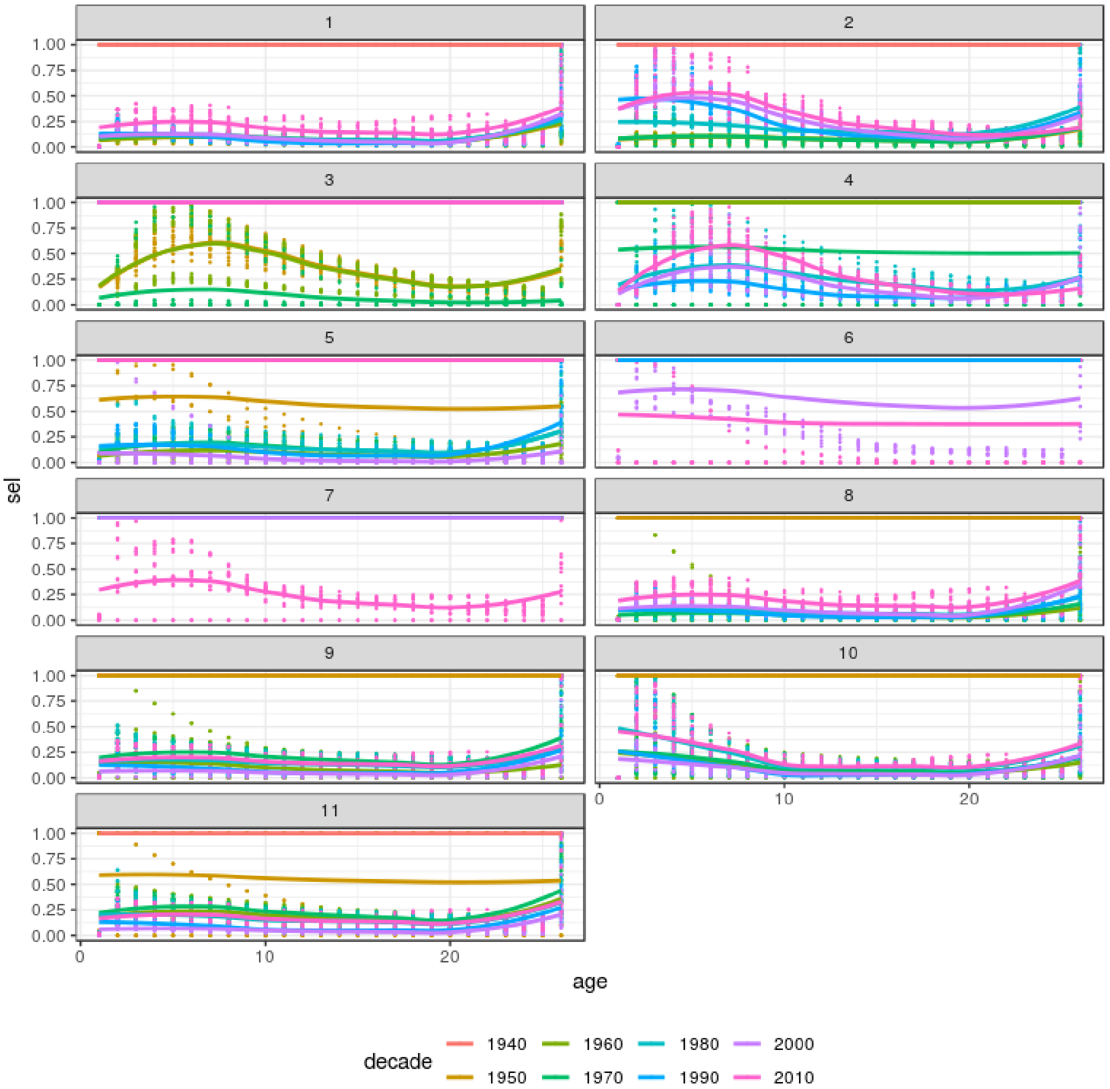


Figure 27. Catch curve analyses from fits.

Residual Analysis

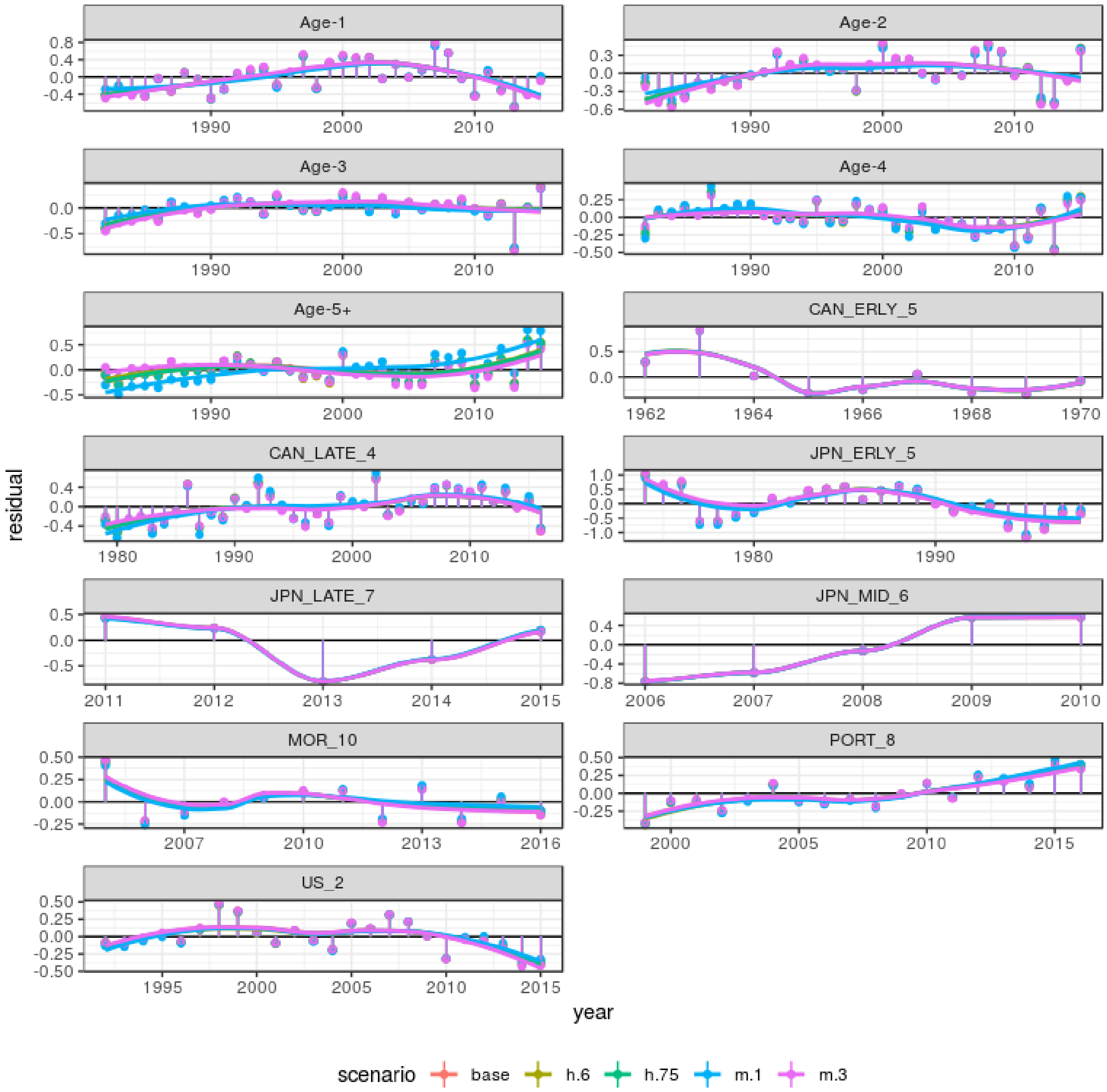


Figure 28. Residuals by year, with lowess smoother.

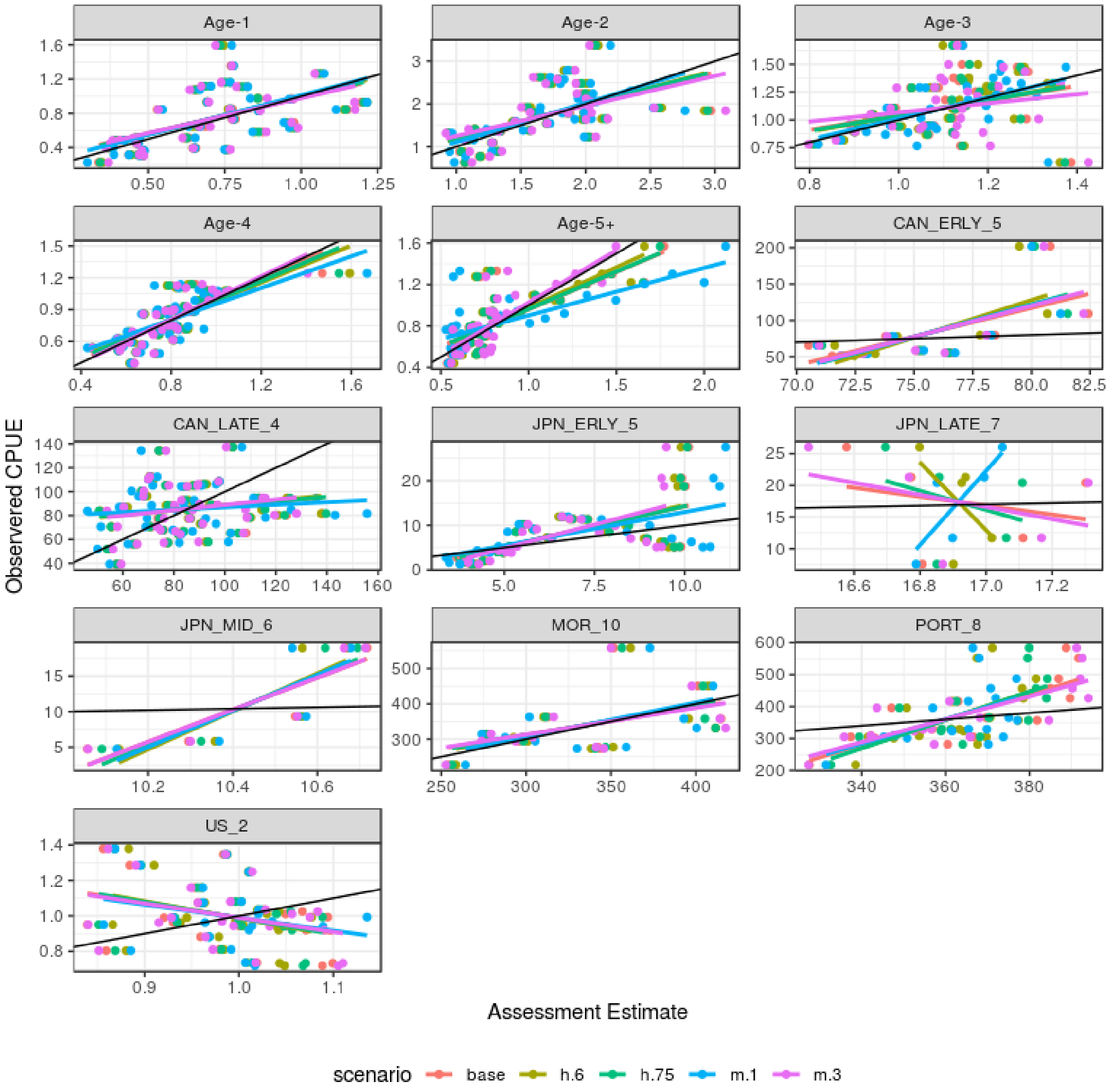


Figure 29. Observed CPUE verses fitted, blue line is a linear regression fitted to the points and the black line is y=x.

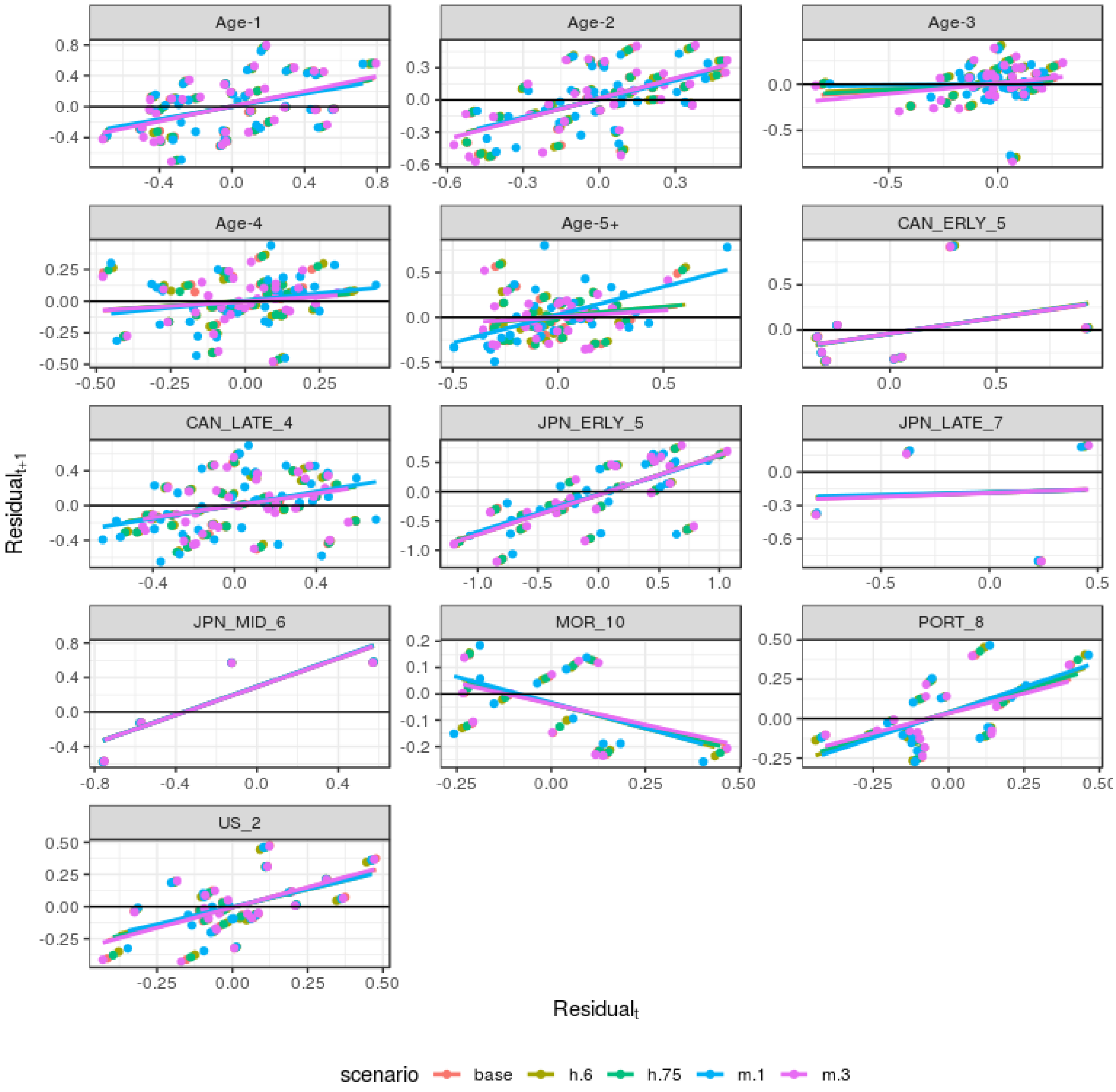


Figure 30. Plot of autocorrelation, i.e. $residual_{t+1}$ verses $residual_t$.

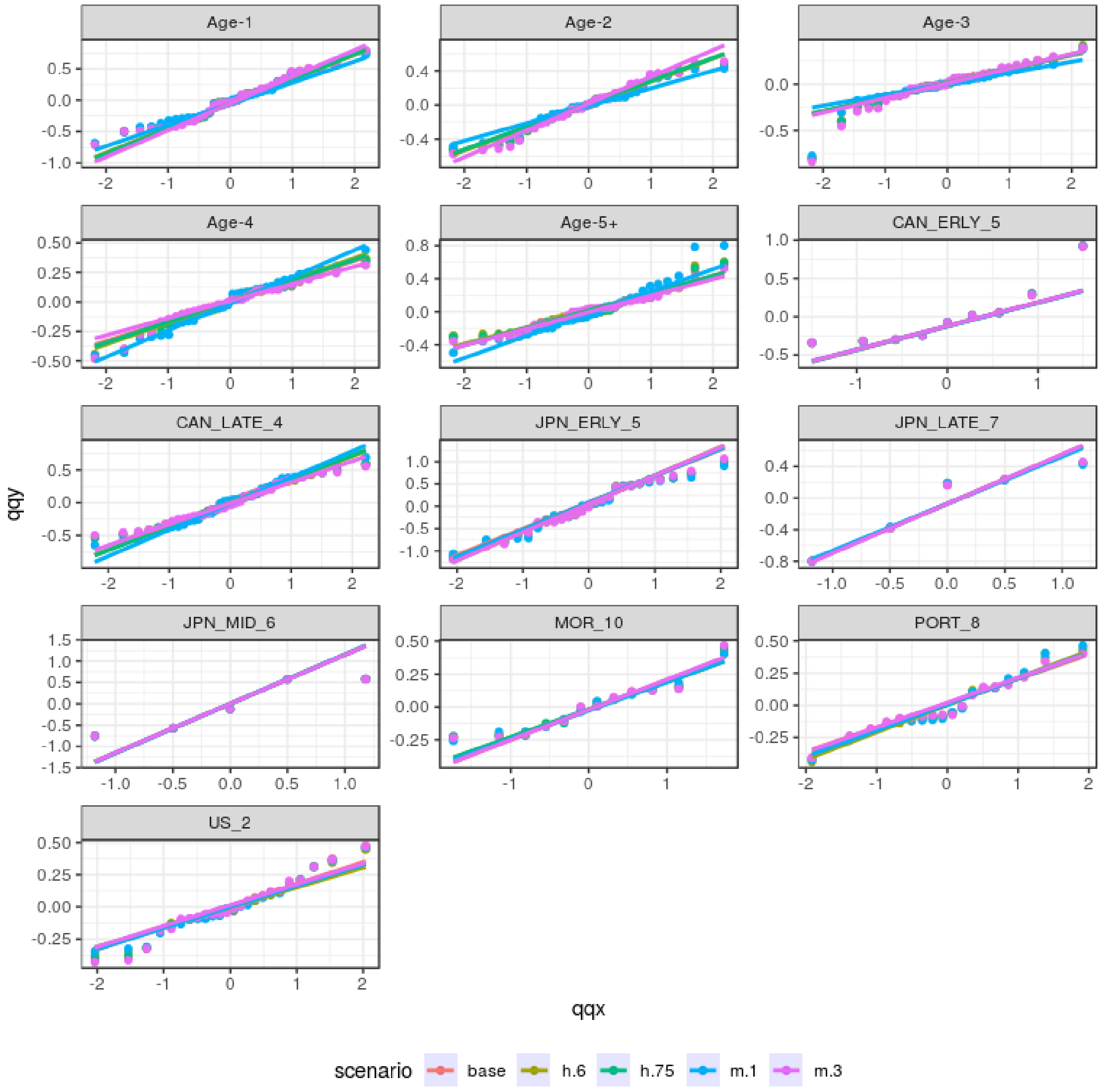


Figure 31. Quantile-quantile plot to compare residual distribution with the normal distribution.

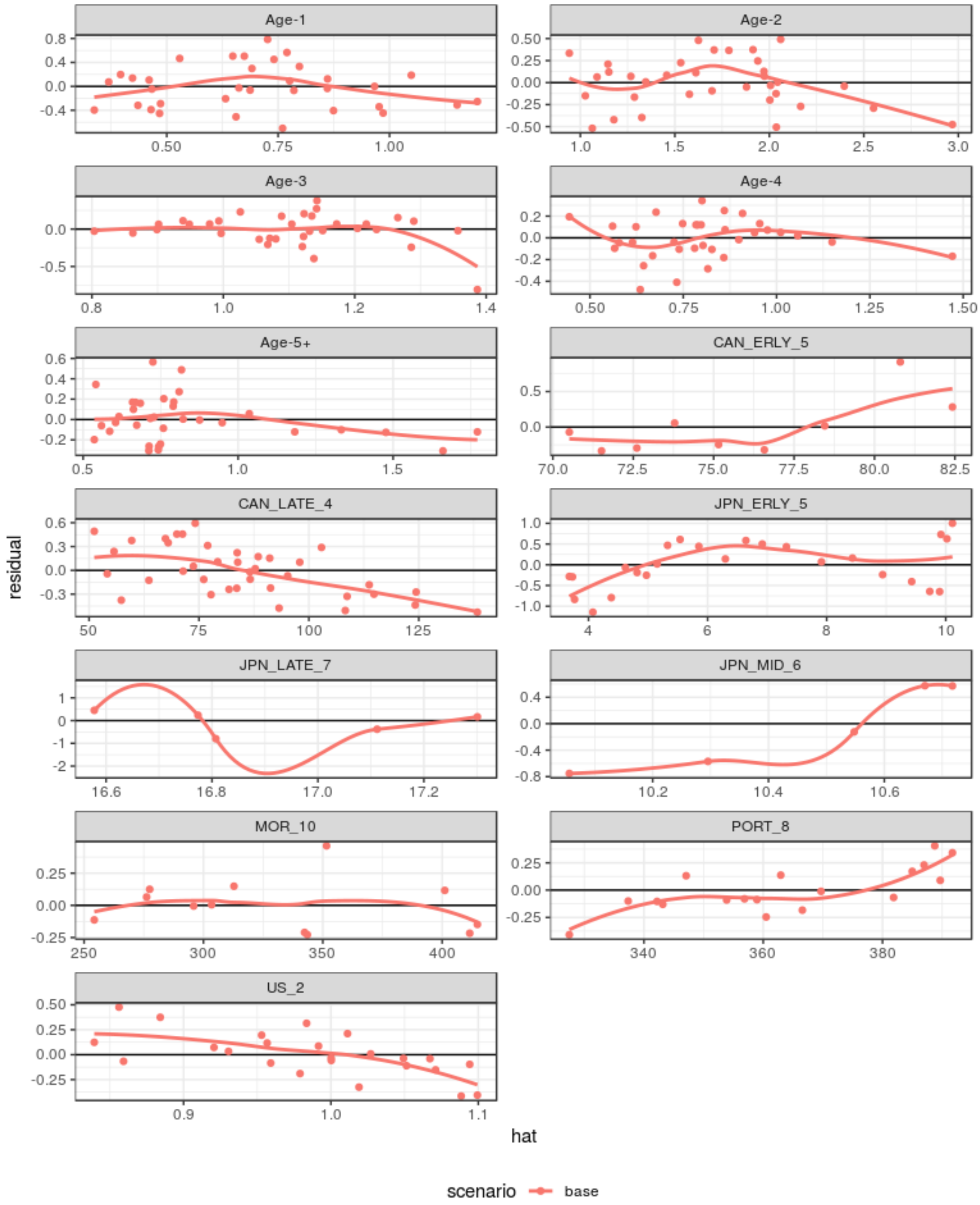


Figure 32. Plot of residuals against fitted value, to check variance relationship.

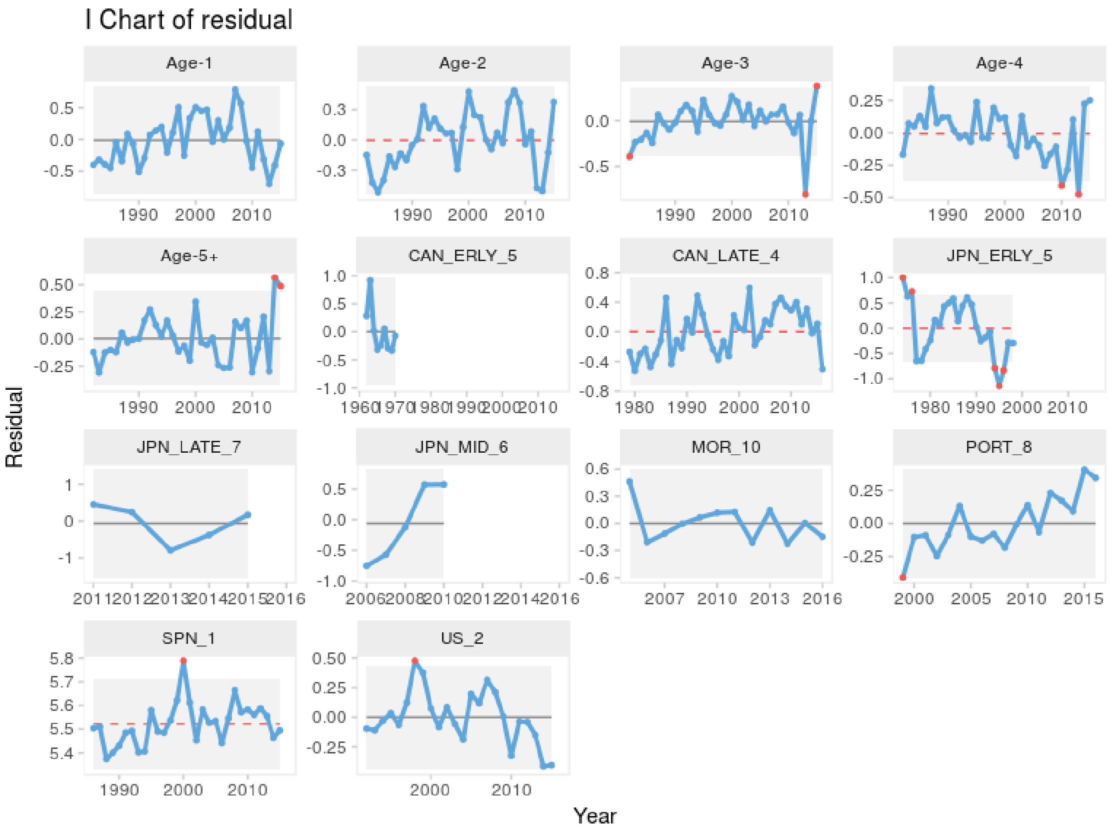


Figure 33. Runs chart for base scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

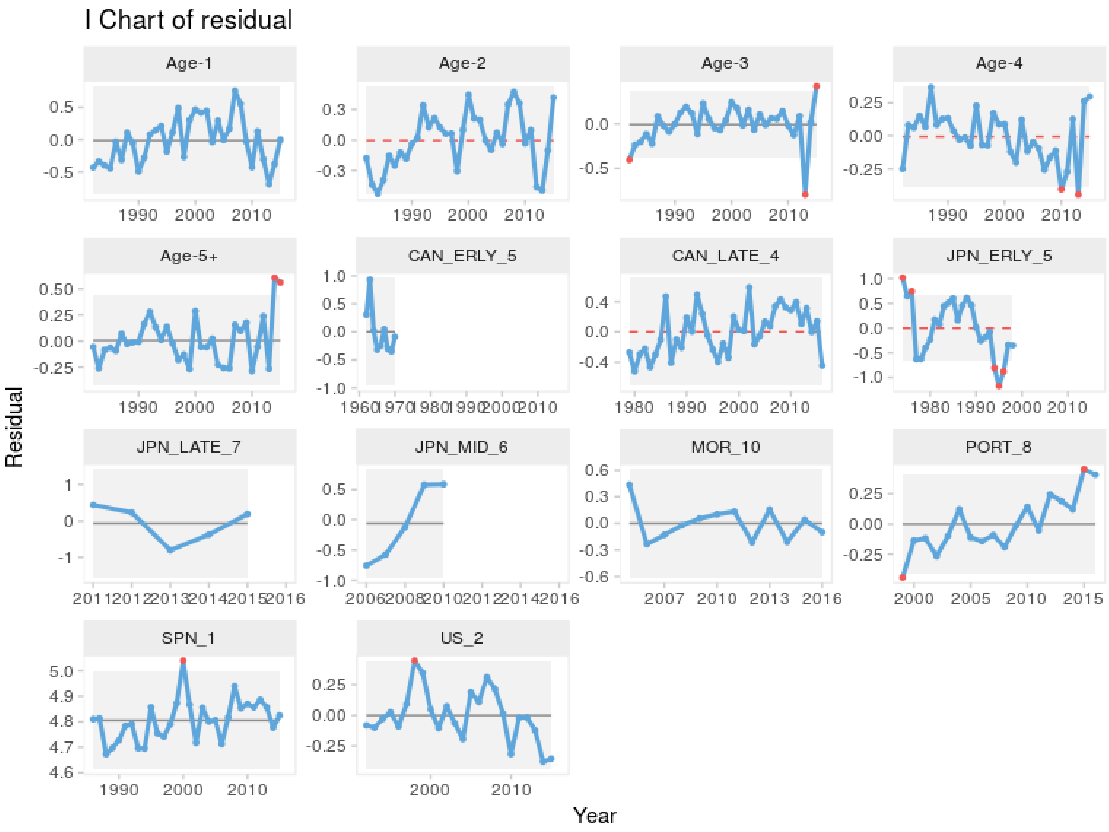


Figure 34. Runs chart for h.6 scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

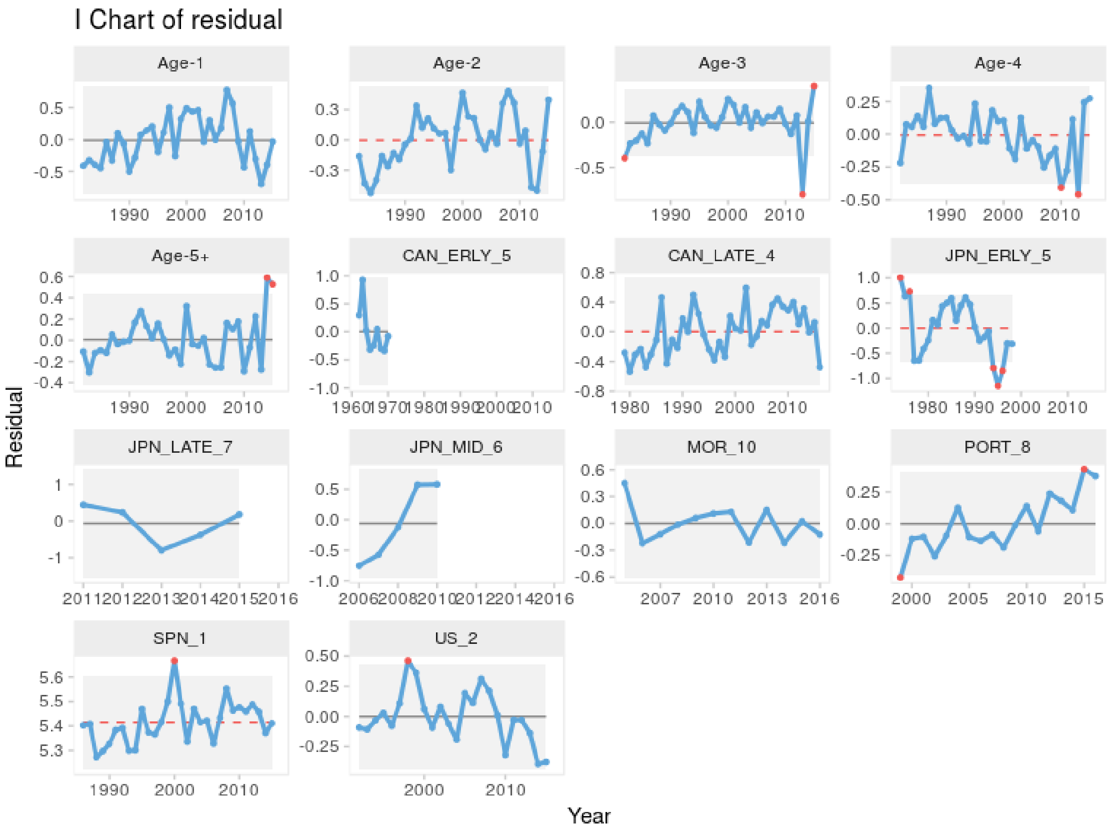


Figure 35. Runs chart for h.75 scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

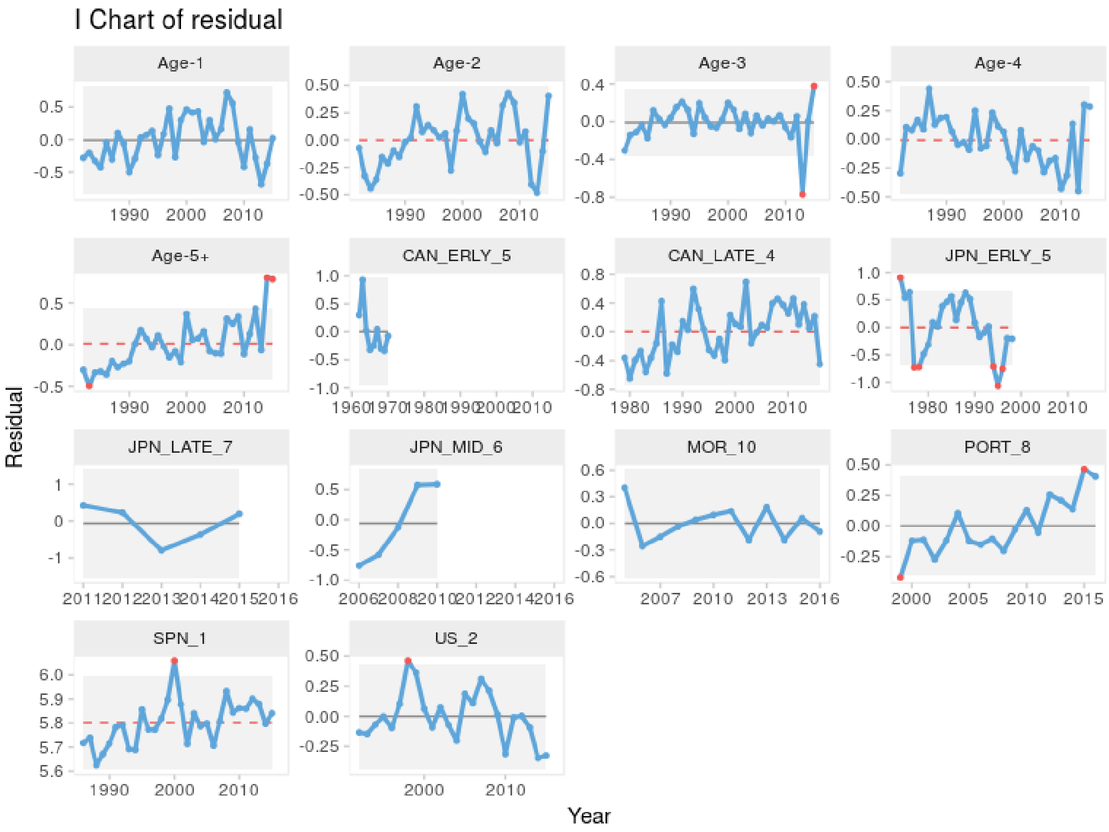


Figure 36. Runs chart for m.1 scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

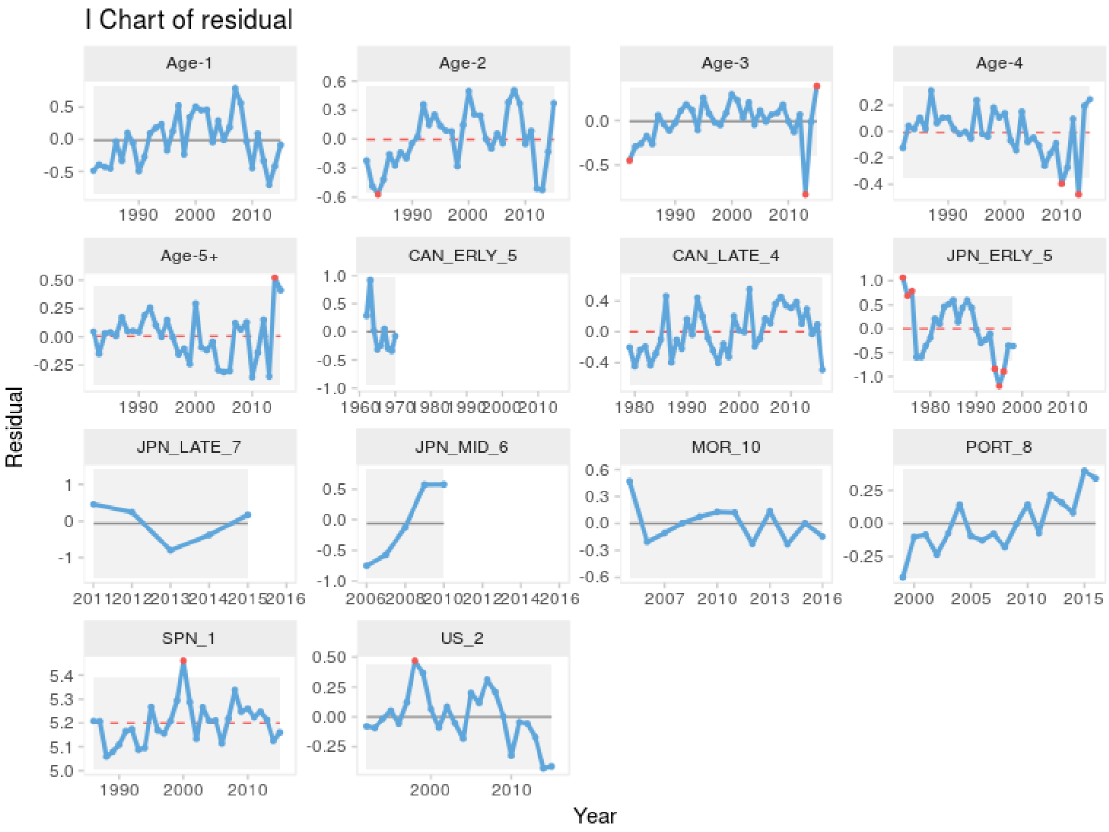


Figure 37. Runs chart for m.3 scenario showing the residuals by year; red points indicate points that violate the 3 sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

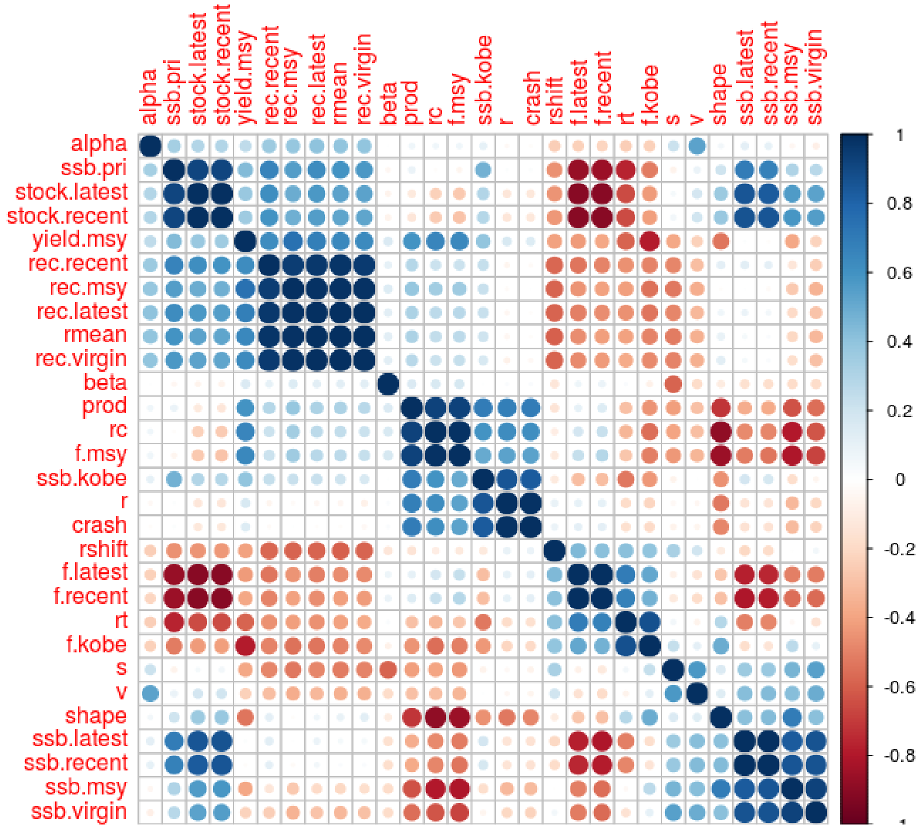


Figure 38. Albacore Indian Ocean correlation matrix.

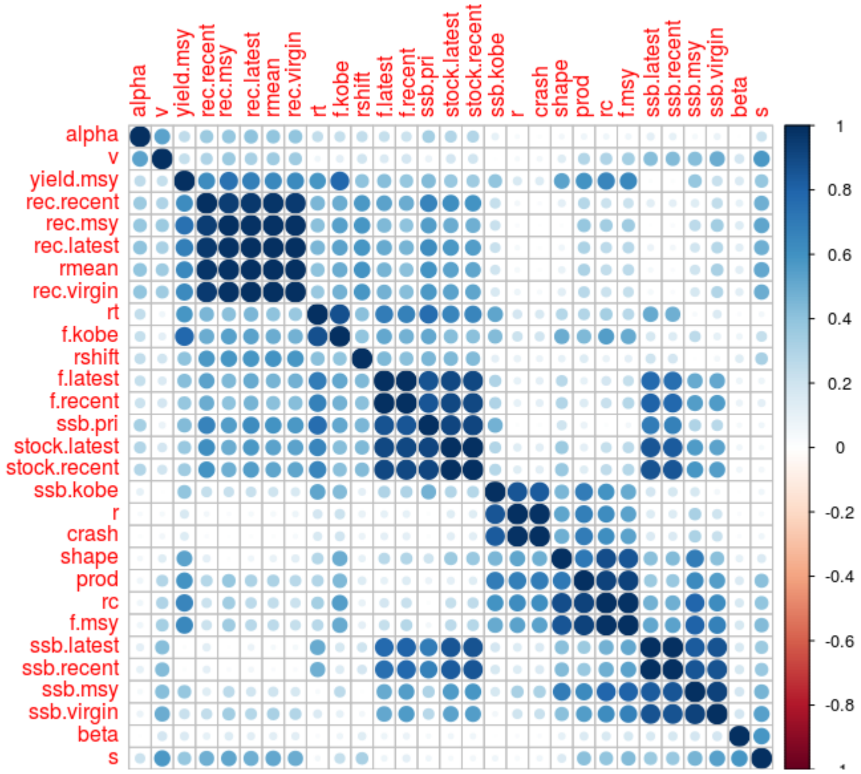


Figure 39. Albacore Indian Ocean correlation matrix, absolute values to show relationships.

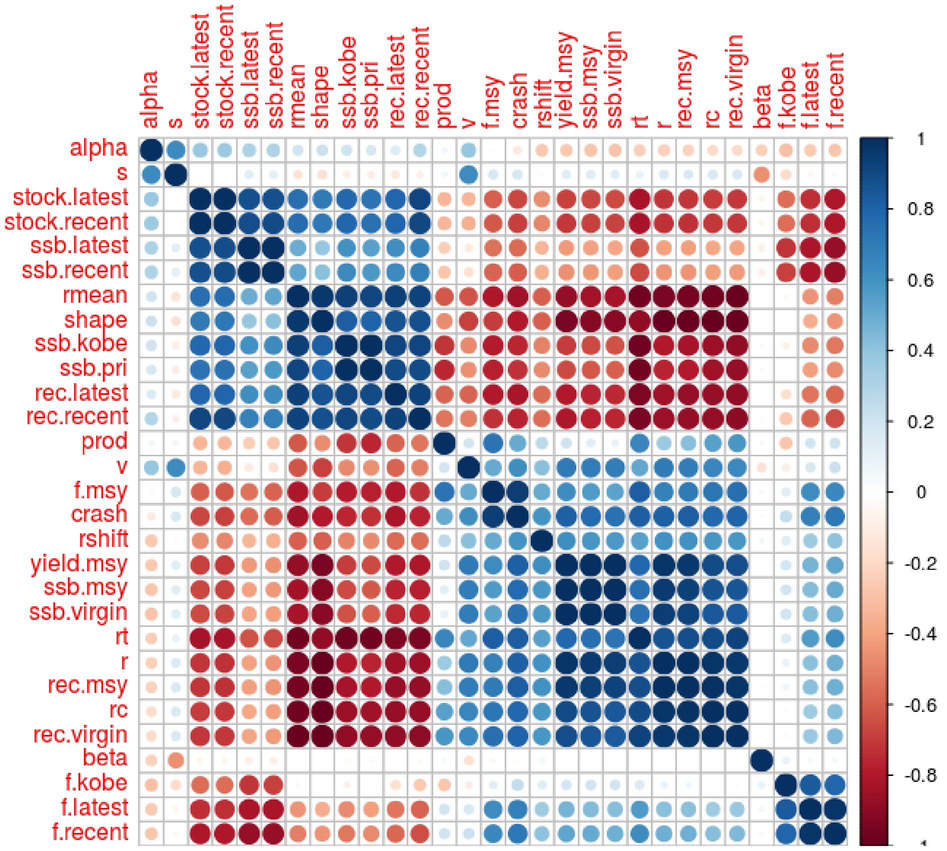


Figure 40. Albacore North Atlantic Ocean correlation matrix.

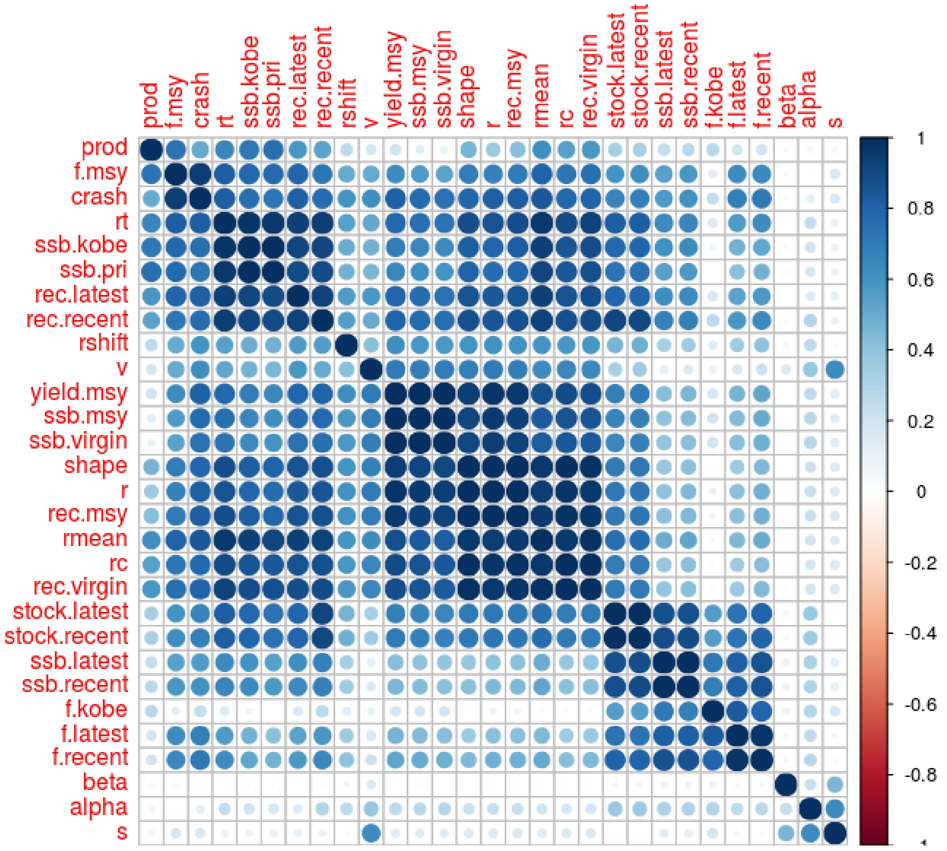


Figure 41. Albacore North Atlantic Ocean correlation matrix, absolute values to show relationships.

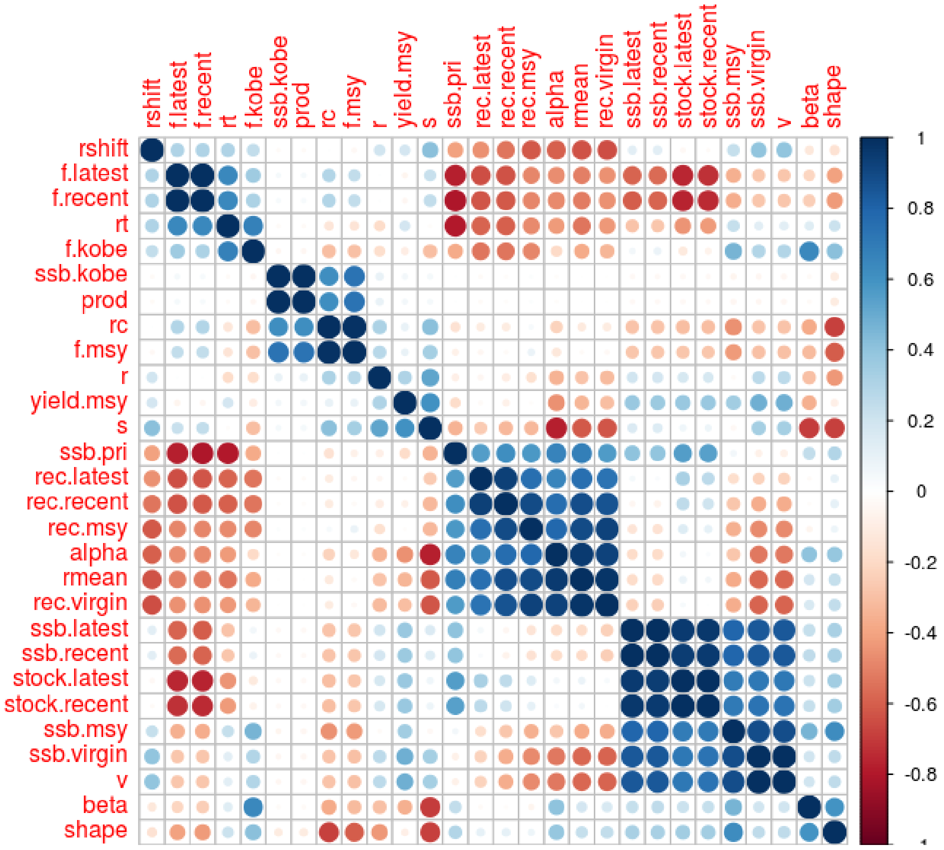


Figure 42. Swordfish Indian Ocean correlation matrix.

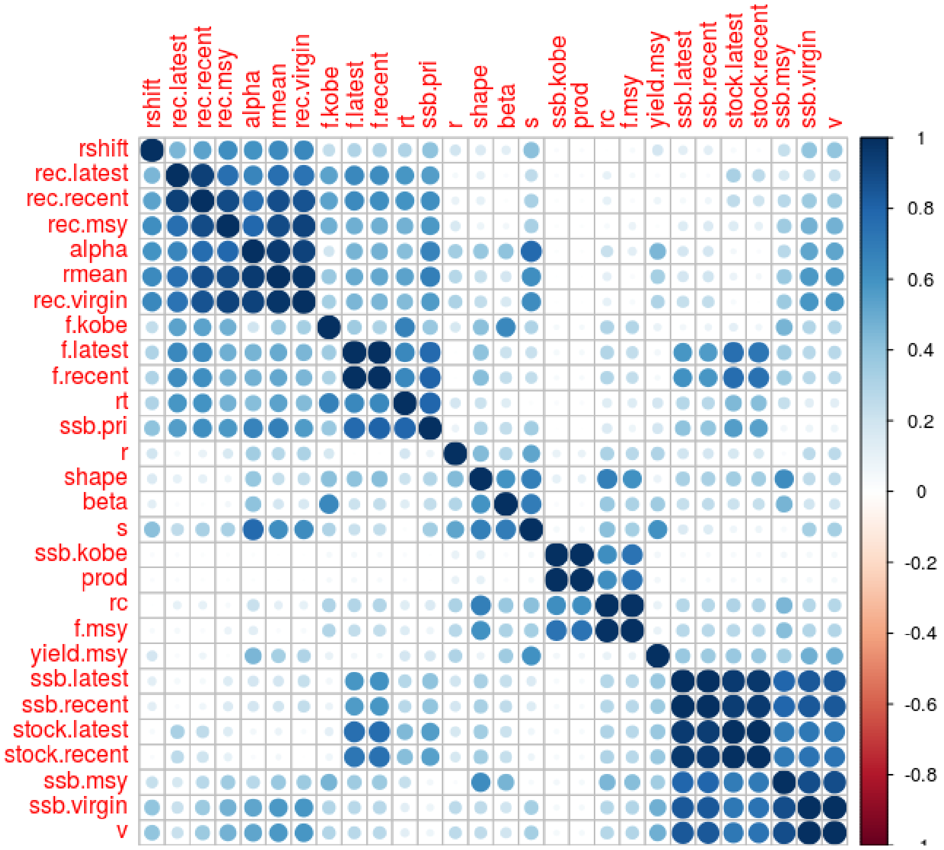


Figure 43. Swordfish Indian Ocean correlation matrix, absolute values to show relationships

Clusters

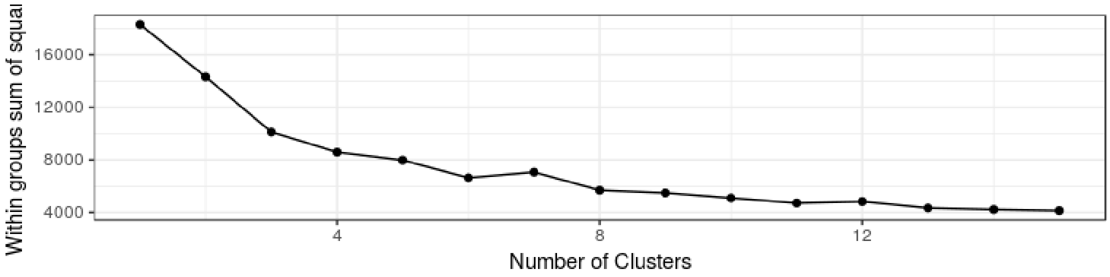


Figure 44. Albacore Indian Ocean cluster selection.

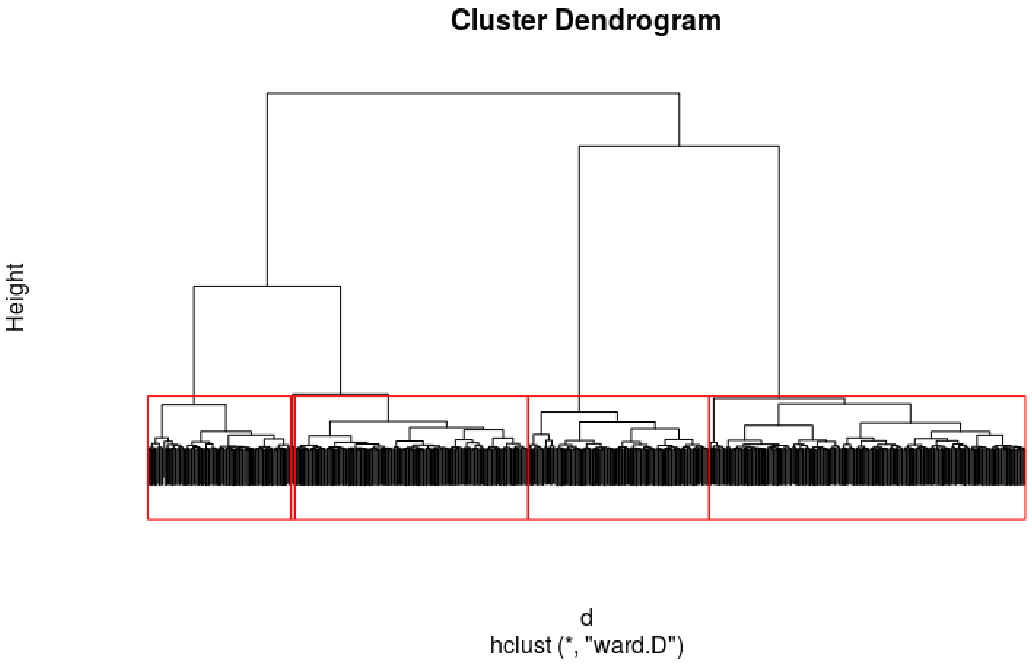


Figure 45. Albacore Indian Ocean Clusters.

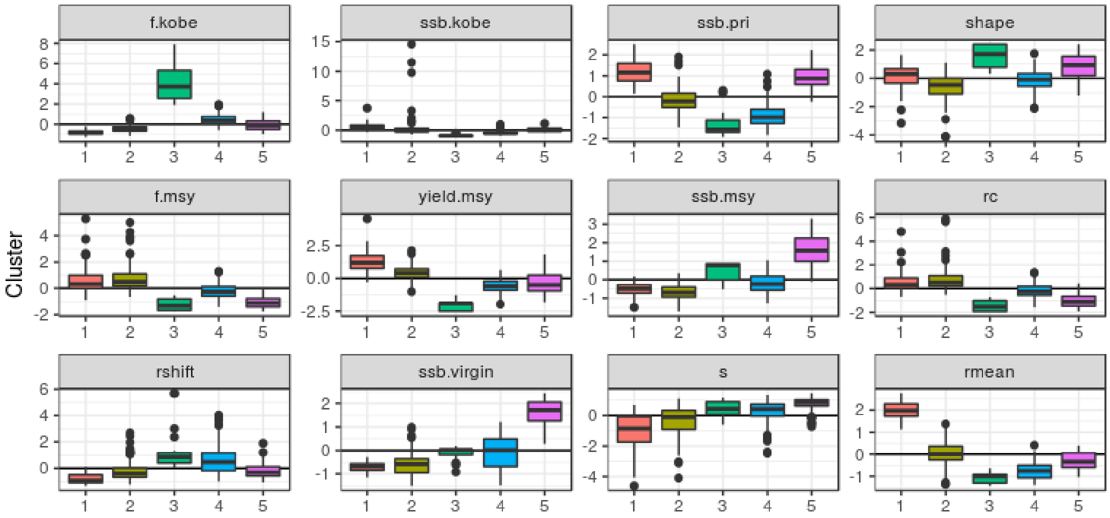


Figure 46. Albacore Indian Ocean Clusters.

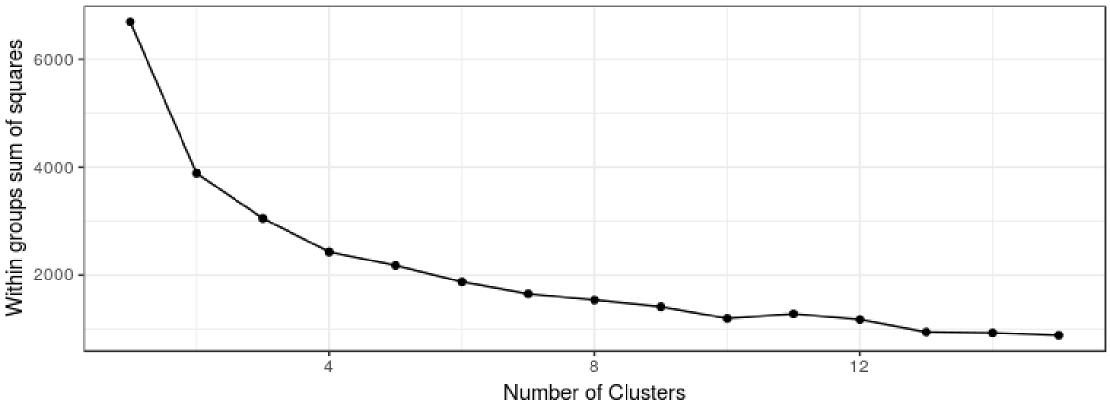


Figure 47. Albacore North Atlantic Ocean cluster selection.

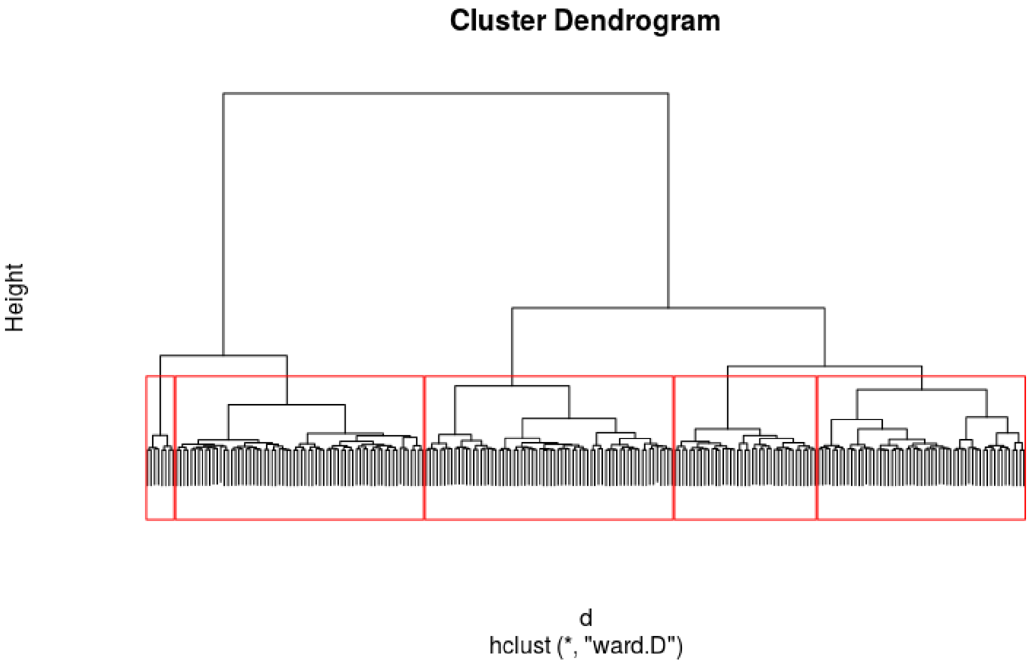


Figure 48 Albacore North Atlantic Ocean Clusters.

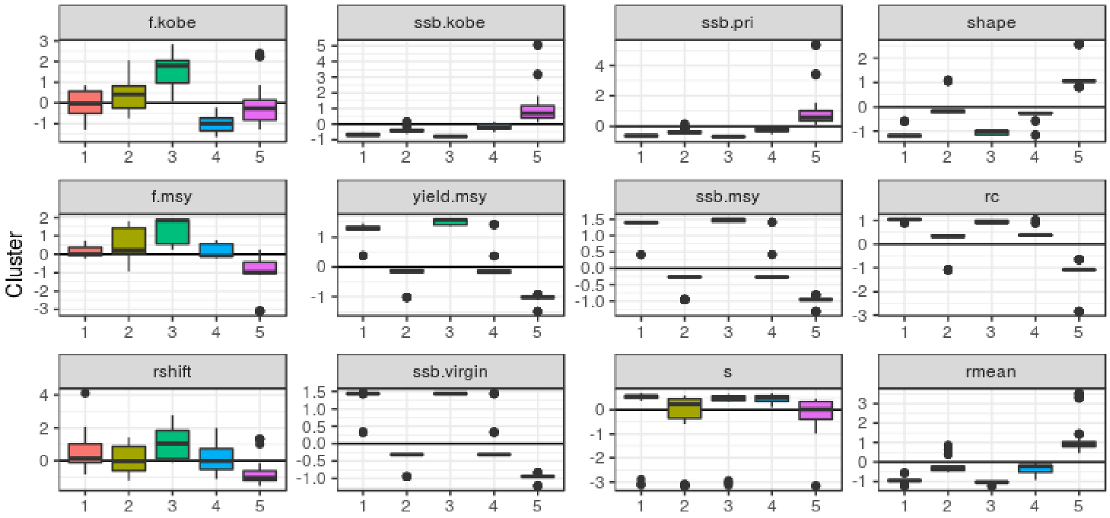


Figure 49. Albacore North Atlantic Ocean Clusters.

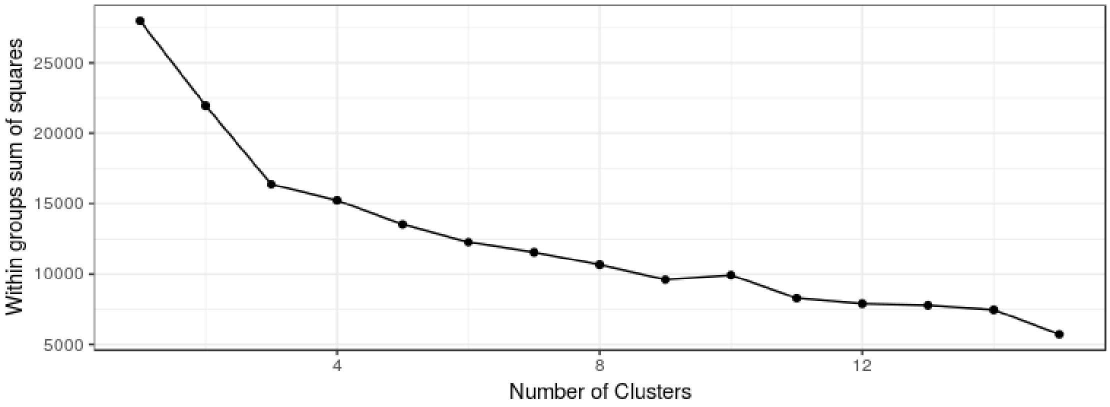


Figure 50. Swordfish Indian Ocean cluster selection.

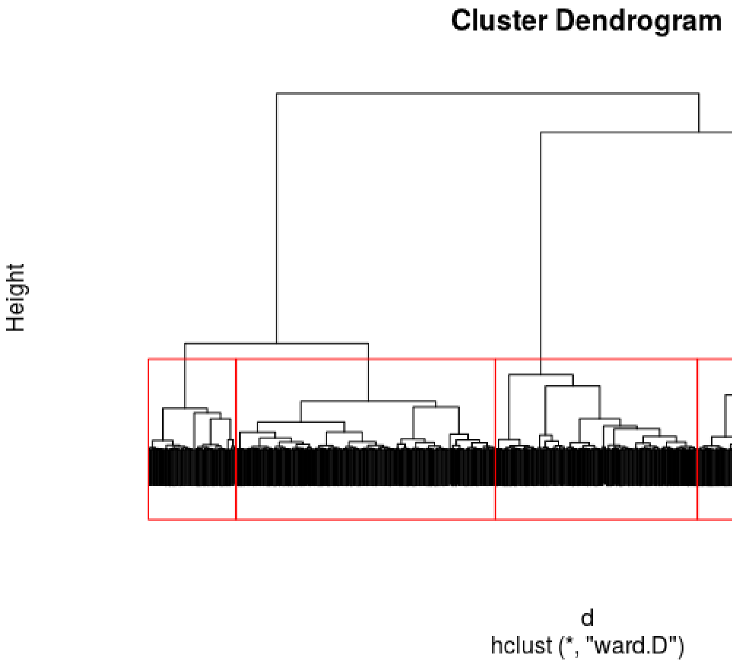


Figure 51. Swordfish Indian Ocean Clusters.

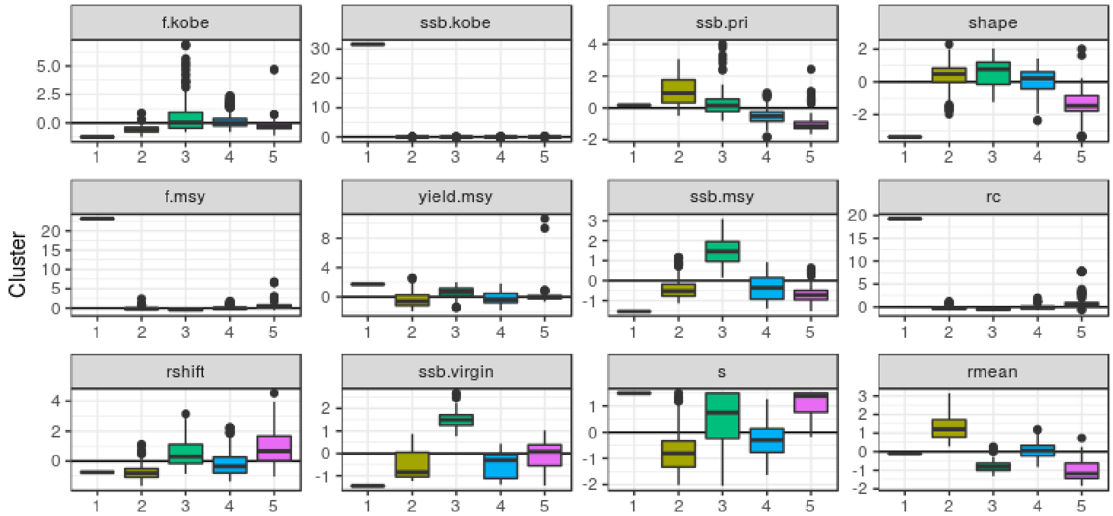


Figure 52. Swordfish Indian Ocean Clusters

Time Series

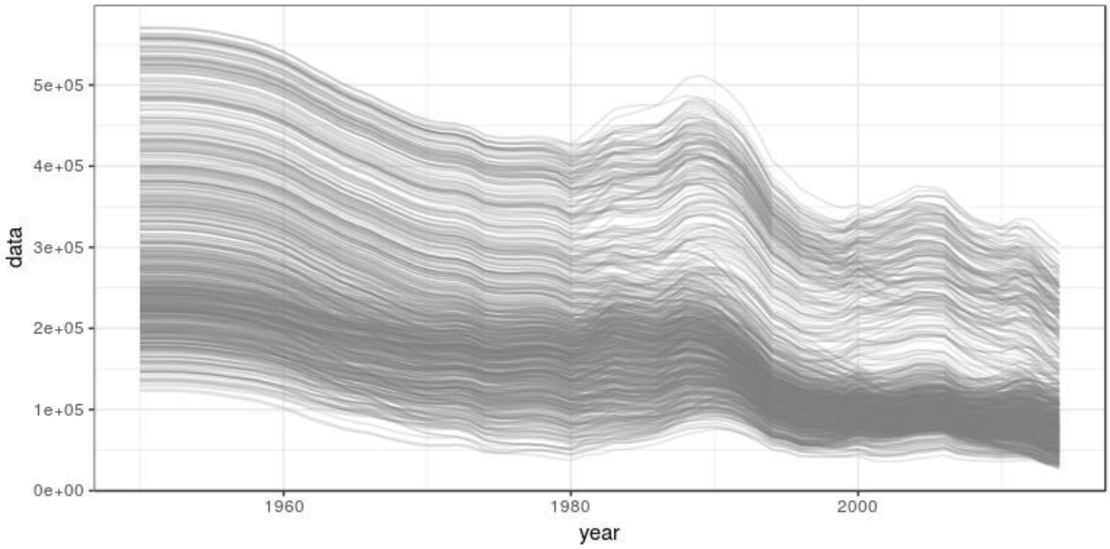


Figure 53. Albacore Indian Ocean SSB.

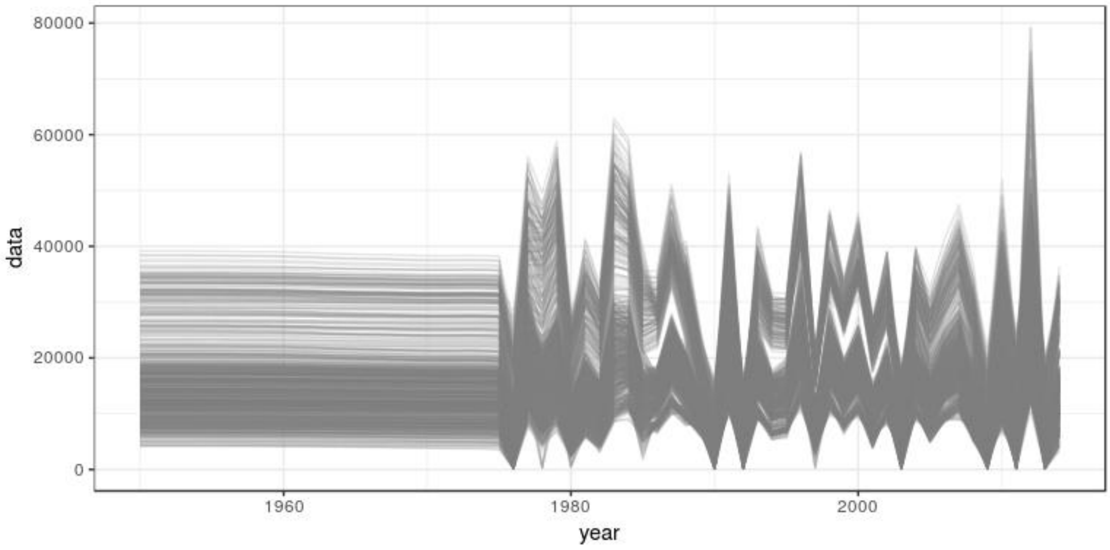


Figure 54. Albacore Indian Ocean recruits.

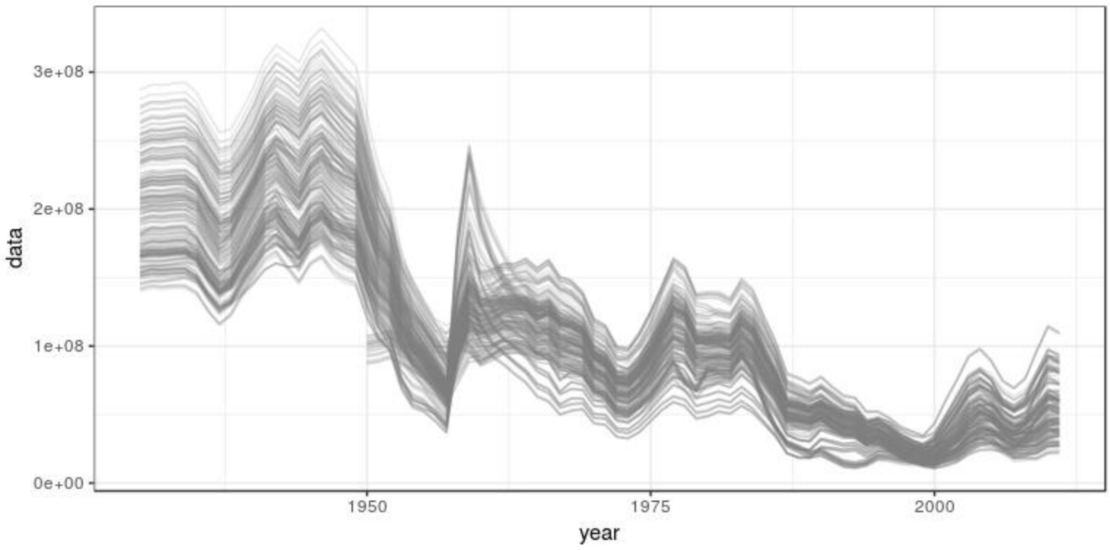


Figure 55. Albacore North Atlantic Ocean SSB.

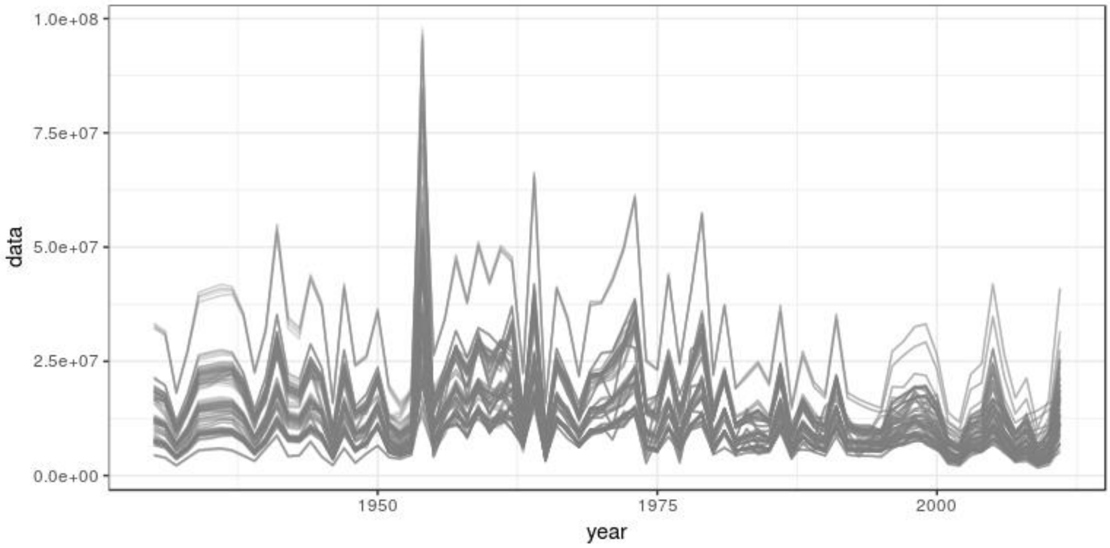


Figure 56. Albacore North Atlantic Ocean recruits.

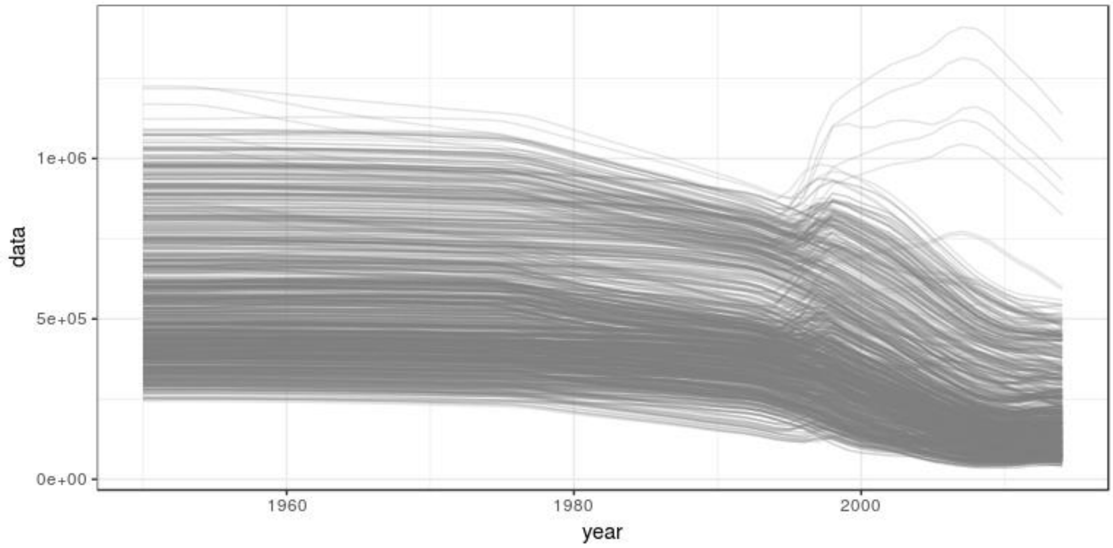


Figure 57. Swordfish Indian Ocean SSB.

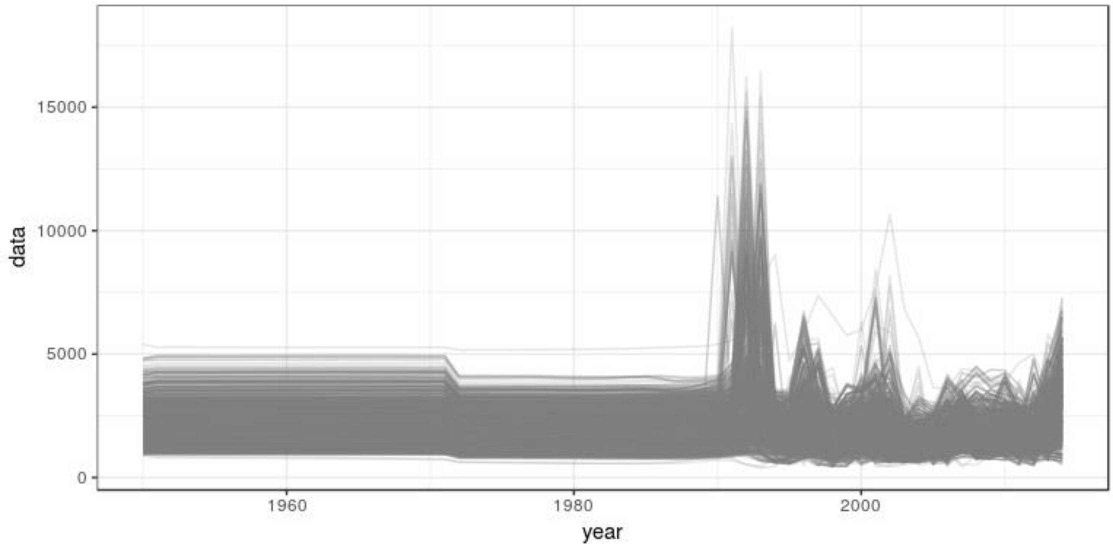


Figure 58. Swordfish Indian Ocean recruits.