A framework for the standardisation of tropical tuna purse seine CPUE: application to the yellowfin tuna in the Indian Ocean.

Katara Isidora[1], Gaertner Daniel[2], Chassot Emmanuel[3], Soto Maria[4], Abascal Francisco[5], Fonteneau Alain[6], Floch Laurent[2], Lopez Jon[7], Cervantes Antonio[8]

[1] Fisheries Consultant, Greece/Portugal; email: isidora10@yahoo.com
[2] IRD, Sète, France
[3] IRD, Victoria, Mahé, Seuchelles
[4] IEO, Madrid, Spain
[5] IEO, Tenerife, Canaries, Spain
[6] External expert, France
[7] AZTI, Pasaïa, Spain
[8] EU DG MARE, Belgium

Abstract

We revised the existing framework for tuna CPUE standardisation in light of the increasing literature that advocates the use of mixed effects models to account for the characteristics of logbook data. We apply the framework on yellowfin tuna (YFT) from the Indian Ocean, caught by the purse seine EU fleet (Spain and France) from 1984 to 2015. We used a comprehensive list of candidate covariates, including non-conventional covariates, and run exploratory models to assess the contribution of each covariate. Due to the large number of covariates, the lasso – least absolute shrinkage and selection operator- method was applied for data mining and model selection purposes. The results are two standardised YFT CPUE time series for the period 1984-2015, one for large fish caught in free-school related sets, and one for mainly juveniles caught in floating object related sets. Issues on the usefulness of highly aggregated data (low resolution: annual and fleet wide) is discussed along with the need for more detailed information on the use of dFADs, preferably at the level of a fishing trip.

Introduction

This paper is the result of the *Workshop for the development of indices of abundance for the EU tropical tuna purse seine fishery*, held in Fuengirola, Spain, on July 2016. During the workshop, experts on tropical tuna fisheries set the foundations for the development of a CPUE standardisation framework for the EU tuna

purse seine fleet. Indian Ocean yellowfin tuna was prioritised as a case study, due to its imminent stock assessment.

The framework is based on the use of generalised linear mixed models, instead of the generalised linear models used to date. CPUE calculation and standardisation primarily uses data from logbooks. The inclusion of random effects in the standardisation model allows us to take into account the special characteristics of fisheries dependent data:

i. Logbook data are longitudinal (Laird and Ware, 1982; Liang and Zeger, 1986), i.e. several observations are made over time on the same subjects, here the subject being the vessel and/or the skipper.
ii. In contrast to the set sampling trajectory of a survey vessel, fishing vessels follow specific strategies and tend to aggregate in areas where fish abundance is high and stable. Thus, areas assumed to be poorer in abundance are not "sampled" by the fishing vessels, and the sampling area varies from one year to the next (Cao et al., 2011).
iii. Closures of Exclusive Economic Zones (EEZs) to the fishery have been observed. Certain EEZs are avoided for other reasons, such as piracy. Hence, the sampling of EEZs is not a random process, but fishing sets are grouped within each EEZ (Fonteneau et al., 2016).
iv. Fishing sets extend over different biogeographical provinces (Longhurst, 2010). We can assume that sets taking place in the same province are not independent observations. In this case the grouping factor is treated as a random effect (Snijders, 2011).

The proposed framework includes the collation of a comprehensive list of new non-conventional variables as potential covariates in the standardisation model; these variables are possible drivers of catchability: fishing strategies technological advances, and environmental factors. The high value on the acquisition of data on covariates on fishing technology at a highly disaggregated scale has already been highlighted by (Bishop, 2006). The covariates constitute factors that affect fishing strategy and variables that describe changes in fishing technology. In the case of the European tropical tuna purse seine fishery

attempts have been made to collate non-conventional information, specifically for dFAD-fishing, and to integrate it into the CPUE standardization procedure (Gaertner et al, 2016). The number of covariates and estimated coefficients can be significant (> 50, as categorical variables may have several levels) and is increasing as new fishing strategies and technologies evolve and more comprehensive data collection protocols are developed. The increasing number of covariates makes model selection a tedious and computationally intensive task. Stepwise selection becomes impossible due to the great number of models one needs to test (for 10 covariates there is 45 different combinations - pairs of covariates - at the 1st step of the forward stepwise regression). Furthermore, stepwise regression cannot deal with collinearity. Statistical procedures of model selection were adapted to the large number of available explanatory variables and the hierarchical structure of the data. For prediction accuracy and interpretation purposes we opt for the lasso – for least absolute shrinkage and selection operator- method for data mining (to discover relationships between CPUE and explanatory variables) and model selection, within a parsimonious approach (Tibshirani, 1996). This approach has been used in the dFAD CPUE standardization of bigeye in the Indian Ocean (Katara et al, 2016). The selection of the final model is based primarily on expert opinion and secondarily on the results derived from the lasso regression and modelling experiments, pertaining to the inclusion/ exclusion of candidate covariates, and the examination of the model estimates and the explained deviance. With these steps we conform to the suggestions given by (Bishop, 2006), and aim for realistic model fits, paying relatively less attention to precision or maximizing explained variance, while adopting modern statistical methods to overcome known issues in fisheries-dependent data.

Material and Methods

The analysis was based on logbook data from two of the main fleets: the French and Spanish purse seiners targeting tuna in the Indian Ocean. The database was subset into 2 datasets: i. free-school sets, which capture large yellowfins and ii. floating object-related sets whose yellowfin component is mainly composed by juveniles. CPUE was calculated and standardised for each dataset. For details on the size frequencies caught per set type see annex I.

The logbook database is managed by the Tuna Observatory of IRD and by the IEO), for the French and Spanish fleets, respectively. Complementary data from other sources were compiled, to be used as covariates in the CPUE standardization models. A full list of the covariates and their sources is given in table 1.

CPUE was defined as the catch per set. Due to the zero-inflated nature of the response variable, a delta log-normal model was applied. The model comprises 2 sub-models:

i. A binomial model that standardises the probability of a positive set, and
ii. A lognormal model that standardises catch per positive set.

The candidate covariates in table 1 were considered fundamental by the working group of experts that took part in the workshop on *the development of indices of abundance for the EU tropical tuna purse seine fishery* in Fuengirola, Spain. However, some variables were collinear or interacting, and overfitting was examined. The model selection exploration began by applying the lasso regression (GLM) in search of a simple model with high goodness of fit, but most importantly a model that accounts for changes in catchability. Variables with regression coefficients greater than 0 are selected and exploratory models are built based upon this original selection.. In cases where the number variables selected in lasso regression was too large to allow for an ordinary least-squares (OLS) regression, a stricter regression coefficient cut off point was chosen. The basis for this choice was to balance the ratio of the number of necessary variables to the number of observations in the data. To test for this we run models with different numbers of variables based on different regression coefficient cut-off points. In conjunction with lasso regression we run some simple models and examined changes in the model fit, attributed to adding or excluding covariates. Such *exploratory* models are recommended when the inclusion of a variable is dubious. The final model was based on the results of these data mining techniques, on the hierarchical structure of the data and on prior knowledge of the fishery and its evolution in time.

From the full list of covariates discussed by the working group in Fuengirola, Spain, some were ignored due to their low resolution and high collinearity. Annual time series of covariates that refer to whole fleets proved problematic, causing rank deficiency. The phenomenon is common when the variance of the predictor is inadequate for model estimation, i.e. there is no signal reflected in the dependent variable

. In the binomial models, searching time was added as an offset.

Lasso regression estimates can be biased; as recommended by (Friedman et al. (2009); Tibshirani (2011, 1996), after model selection, the final models were estimated using OLS. Predictions were made with lsmeans (Lenth, 2014). Memory leakage issues were dealt with by manually constructing the data reference grid (an array of factor and predictor levels, as described in Lenth, 2014) upon which predictions in lsmeans are based. Confidence intervals were calculated with the delta method (Sobel, 1982; Casella and Berger, 2002). The analysis was performed in R.

Results

Sets on Free Schools

The lasso GLM for the probability of a positive free school set, given searching time, gave the following list of explanatory variables with coefficients > 0.01: year, month, vessel age, EEZ, grid cell, vessel ID, biogeographical province, fishing time, the interaction between year and month, and the interaction between year and grid cell. As mentioned previously, Lasso GLMs narrowed down the number of possible predictors for the final model. Some of the selected variables, e.g. the vessel ID and the interaction between year and grid cell, were considered as random effects, since they relate to the longitudinal structure and the sampling of the data. Fishing time and searching time being correlated, only the second one was used as an offset in the model. The interaction between year and month was also excluded as it caused convergence errors, because of the large number of regression coefficients to be estimated, given the number of available observations. Models with different random effects were also tested and the final model included year,

month, vessel age, and grid cell. The lsmeans predictions for the probability of a positive (catch > 0) free-school set, derived from the final standardisation model are shown in Fig. 1.

The lasso GLM for the YFT catch per positive free school-related set gave the following list of variables with coefficients > 0.01: vessel length, vessel capacity class, ratio of free schools sets vs floating object related sets, fishing time, vessel ID, EEZ, the interaction between year and month, and the interaction between year and grid cell. Based on exploratory mixed models (Figs 2 and 3) and our knowledge of the data structure and derivation, vessel ID, the interaction between year and grid cell, EEZ, and grid cell were included as random, instead of fixed, effects: The standardised YFT CPUE for free-school sets (that is to say, the product of the two sub-models) is shown in Fig. 4.

Sets on floating objects (dFADs and logs)

The lasso GLM for the probability of a positive floating object-related set, given searching time, gave the following list of variables with coefficients > 0.0001: year, month, grid cell, vessel ID, EEZ, time at sea, the ratio of free school to floating object related sets, the interactions between year and month, and between year and grid cell. As with free school sets, we applied different exploratory random effect models and assessed their fit to conclude to the final model with: (i) year, month, grid cell, and the ratio of free school to floating object related sets as fixed effects, and (ii). vessel ID, EEZ, and the interaction between year and grid cell as random effects. Time at sea was not included due to its correlation to searching time.

The lasso GLM for the YFT catch per positive floating object-related set, given searching time, gave the following list of variables with coefficients > 0.01: year, month, grid cell, vessel ID, biogeographical province, EEZ, time at sea, fishing time, searching time, the ratio of free school to floating object related sets, the ratio of YFT to skipjack price, the number of supply vessels, the proportion of BSE type dFADs (dFADs equipped with satellite GPS and ecosounder), the interactions of year and month and of year and grid cell. Of these variables, 3 were automatically dropped from the GLMMs (rank deficiency); these were the ratio of YFT to skipjack price, the number of supply vessels, and the proportion of BSE type FADs. The biogeographical provinces were also excluded because the contribution of the factor was low (exploratory models and

BIC).The final selection of fixed effects consists of year, month, fishing time, the ratio of free school to floating object related sets, and the grid cell. The following were included as random effects: vessel ID, the interaction of year and EEZ, and the interaction of year and grid cell.

The lsmeans predictions for the two sub-models, and for the standardised CPUE for floating object related sets (as the product of the two sub-models) are presented in Fig. 5.

Discussion

Two main issues arise from the standardisation exercise presented in this paper: the need for accurate and high resolution covariates and the need for a framework that takes into account the spatial structure of fishing effort and the hierarchical structure of the data.

Good quality of covariates refers to the relevance, the accuracy, and the resolution of the covariates. For example the number of supply vessels given at an annual and fleet wide scale, has a low resolution and is discarded from the models. However it is well documented that the use of supply vessels increases the capacity of the fishing vessels. A more relevant and informative covariate could refer to than the number of times the fishing vessel used a supply vessel per fishing trip. Similarly, the use of dFADs and the related strategies, developed by the skippers, are fairly complex; they cannot only be captured by an annual time series of the number of dFADs in the Indian Ocean. We need to understand the use of dFADs in depth to be able to choose the right covariates. Our understanding to date dictates that disaggregated information, possibly at the level of the fishing trip or fishing day, is needed for dFADs related covariates to be informative in the framework of CPUE standardisation. Such annual time series or trends could not be used in CPUE standardisation models, in the current study because they were collinear – as trends tend to be due to their underlying temporal autocorrelation. Their resolution is low and they were automatically dropped from the models or they were selected out during model selection procedures, as non-informative. For small subsets of the fleet, where information at a set level was available (e.g. the distance of the set from the main dFADs area, the number of dFADs in the vicinity of the set) the covariates relating to dFADs were successfully

included in the CPUE standardization models (e.g., Gaertner et al, 2016 for the results of the EU research project CECOFAD on these aspects). Considering that the tuna targeting purse seine fleets are a unique source of information on juvenile YFT, a data call on dFADs usage for the EU fleet seems imperative.

Despite some data deficiencies this paper sets the basis for a new CPUE standardisation framework, applied for the first time on a combination of logbook data and non-conventional variables from the EU fleet (Spanish and French). A review of the CPUE standardisation framework was needed in light of new statistical techniques and the increasing literature on the advantages of mixed modelling. Indeed, mixed models prove useful for the standardisation of CPUE because they allow us to analyse longitudinal data, and to account for impacts of large spatial heterogeneity of fishing efforts and for spatial/temporal dependencies. Mixed modelling is an active research field for statisticians and fisheries scientists alike. As algorithms improve we will be able to overcome computational hitches and develop models that fully capture the complexity of the data.

Acknowledgements

The authors would like to thank all the participants and observers in the workshop for *the development of indices of abundance for the EU tropical tuna purse seine fishery*, held in IEO (Fuengirola, Spain), on July 2016, and all the experts that offered their feedback. A special thanks to Mauricio Ortiz for his advice and helpful remarks. The workshop and the study, on which the current paper is based, was supported by EU DG MARE.

References

Bishop, J., 2006. Standardizing fishery-dependent catch and effort data in complex fisheries with technology change, J. Rev. Fish Biol Fisheries 16, 21–38.

Cao, J., Chen, X., Chen, Y., Liu, B., Ma, J., Li, S., 2011. Generalized linear Bayesian models for standardizing CPUE: an application to a squid-jigging fishery in the northwest Pacific Ocean. Sci. Mar. 75, 679–689.

Casella, G., and Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.

Fonteneau A., D. Gaertner, and P.J.P. Alayón (2016) An overview of detailed CPUEs and of fishery indicators of the EU purse seiners in the Atlantic. Document SCRS/2016/183.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent . J. Stat. Softw. 33, 1-22.

Gaertner D, Ariz J, Bez N, Clermidy S, Moreno G, Murua H, Soto M (2016). Results achieved within the framework of the EU research project: Catch, Effort, and eCOsystem impacts of FAD-fishing (CECOFAD). SCRS/2016/030

Katara I, Gaertner D, Maufroy A, Chassot E. (2016) Standardization of catch rates for the Eastern tropical Atlantic bigeye tuna caught by the French purse seine FAD fishery. Collect. Vol. Sci. Pap. ICCAT, 72(2): 406-414

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Lenth, R., 2016. Least-Squares Means: The R Package lsmeans. J. Stat. Softw. 69, 1-33 Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Longhurst, A.R., 2010. Ecological geography of the sea. Academic Press.

Snijders, T.A., 2011. Multilevel analysis. Springer.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological methodology, 13, 290-312.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 273–282.

Tables

Table 1 List of candidate available covariates for CPUE standardisation models.

Covariate	Unit / format	Type	Source
Year of fishing operation	year	factor	Derived from date of activity, i.e. d_act of the table ACTIVITY of BALBAYA
Month of operation	month	factor	Derived from date of activity, i.e. d_act of the table ACTIVITY of BALBAYA
5x5 grid square / cell	CWP	factor	Sheet CWP grid; http://www.fao.org/fishery/cwp/en
Unique vessel identifier		factor	c_quille from TURBOBAT
Fleet segment: FRA from French and associated flags (ITA, MYT, etc.) and ESP for Spanish and associated flags (SYC, etc.)	FRA or ESP	factor	Derived from 'pays' from TUBOBAT
Age of the vessel at the time of operation	years	numeric	year of activity – initial year of service (an_serv from TURBOBAT)
Length overall of the vessel	M	numeric	v_l_ht from TURBOBAT
Storage capacity of the vessel	m3	numeric	v_ct_m3 from TURBOBAT
Capacity class of the vessel in metric tons	8 classes	factor	c_cat_b from TURBOBAT
Vessel class of capacity in metric tons	8 classes	string	l_capac from table CAT_BATEAU of BALBAYA
Cumulated time at sea spent by the vessel in the stratum	hour	numeric	v_tmer from table ACTIVITE of BALBAYA
Cumulated fishing time spent by the vessel in the stratum	hour	numeric	v_tpec from table ACTIVITE of BALBAYA
Cumulated searching time spent by the vessel in the stratum	hour	numeric	(v_tpec – v_dur_cal) from table ACTIVITE of BALBAYA
Cumulated number of fishing sets by the vessel in the stratum		numeric	v_nb_calees from table ACTIVITE of BALBAYA
Cumulated number of successful fishing sets by the vessel in the stratum		numeric	v_nb_calee_pos from table ACTIVITE of BALBAYA
Longhurst province of origin of the catch	ProvCode	factor	Shape file Longhurst_world_v4_2010
Exclusive Economic Zone of origin of the catch	ISO_3digit	factor	Shape file VLIZ_EEZ; ABNJ = Areas Beyond National Jurisdiction
Cumulated number of fishing sets divided by the cumulated number of fishing days spent in the stratum		numeric	
Cumulated number of fishing sets divided by the cumulated number of searching days spent in the stratum		numeric	
Annual Proportion of type of buoys by flag and ocean		numeric	
Annual Total number of FADs (Atlantic only, but trend supposed to be similar in Indian Ocean)		numeric	
Annual Total number of support vessels (Indian Ocean only)		numeric	
Monthly Prices by commercial category		numeric	Bangkok's market
Mixed Layer Depth [5° square*month]		numeric	NCEP GODAS

Figures

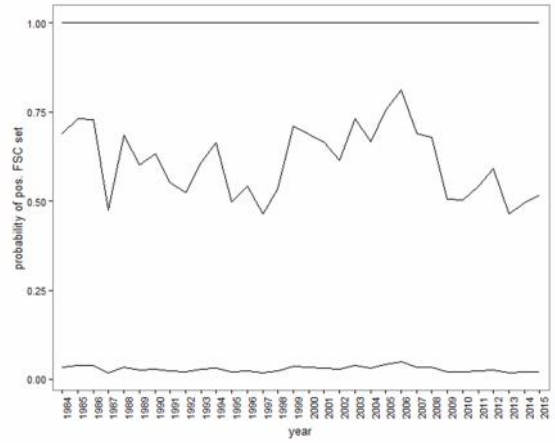


Figure 1 lsmeans predictions for the probability of a positive (catch > 0) free-school set, with 95% CIs.

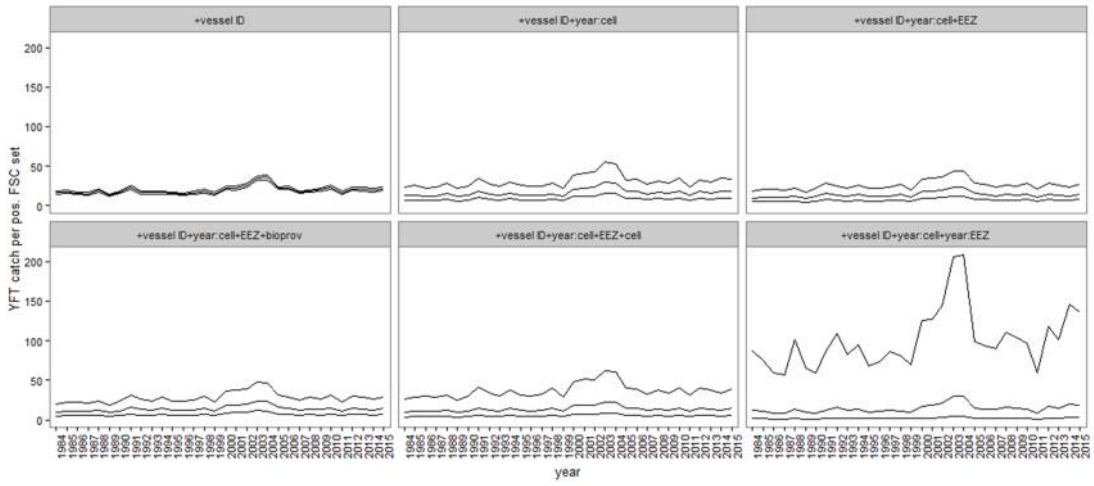


Figure 2 How does adding random effects change the Confidence Intervals of the predicted values. The graphs show the predictions for YFT catch per positive set. The fixed effects part of the model consists only of 1 predictor, year. We start by adding vessel ID as a random effect. The CIs change as more random effects are added. The results of such exploratory models (graphical inspection and BIC) serve as a guide for the inclusion/exclusion of random effects. Here the interation of EEZ with year results in a big increase of BIC and large confidence intervals, possibly due to overfitting. Also EEZs that were not sampled on specific years lie in the boundary of the fishing area and have already been screened out during the data cleaning; therefore, the inclusion of this interaction is not necessary. The inclusion of the grid cell as a random effect is borderline (decrease in BIC < 10, visible effect on CIs) and needs to be revisited in the context of a full model.

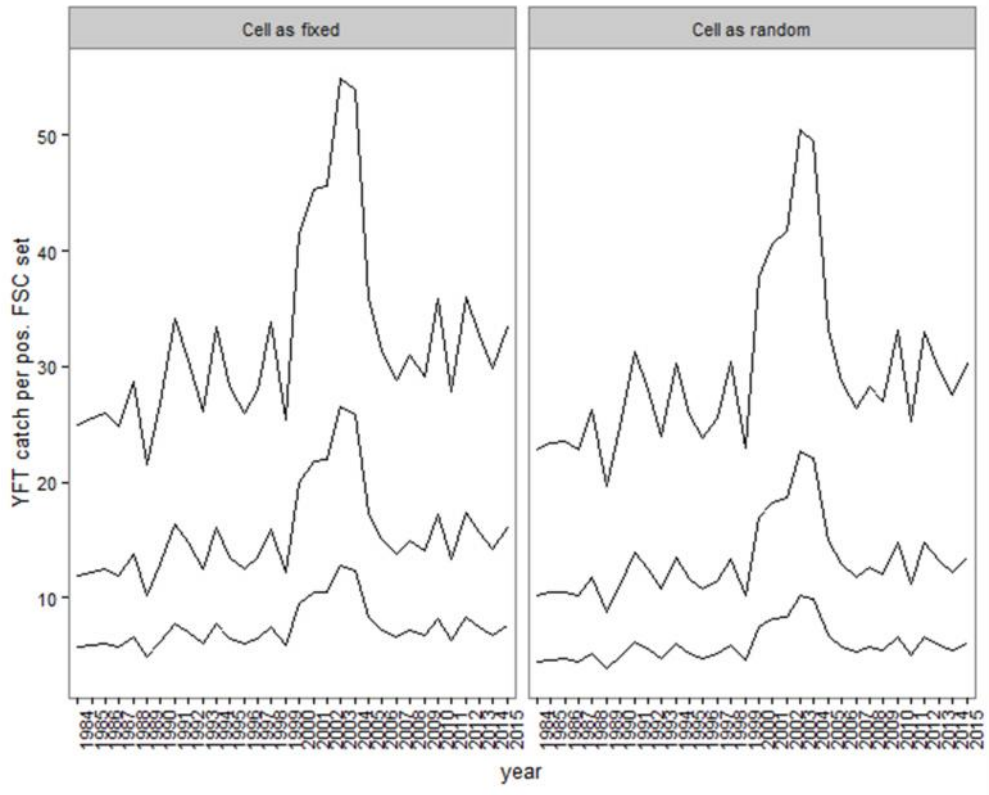


Figure 3 Using grid cell as a random or fixed effect does not have a great impact on the trend or the 95% CIs. The BIC for the model that treats grid cell as random effects is lower.

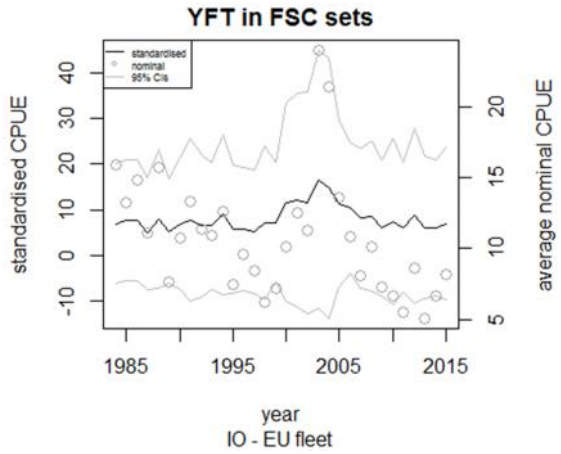


Figure 4 shows the standardised YFT CPUE (the product of the two sub-models) for free-school sets. The confidence intervals are very wide, highlighting the high degree of dependence between observations at different levels of the data structure.

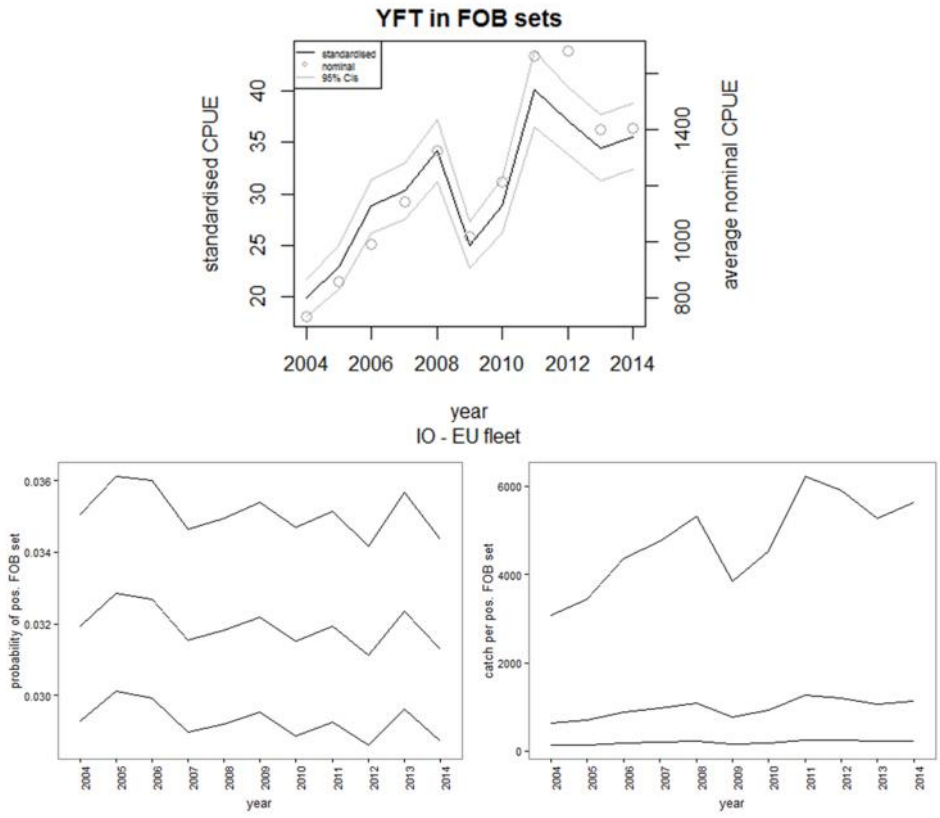
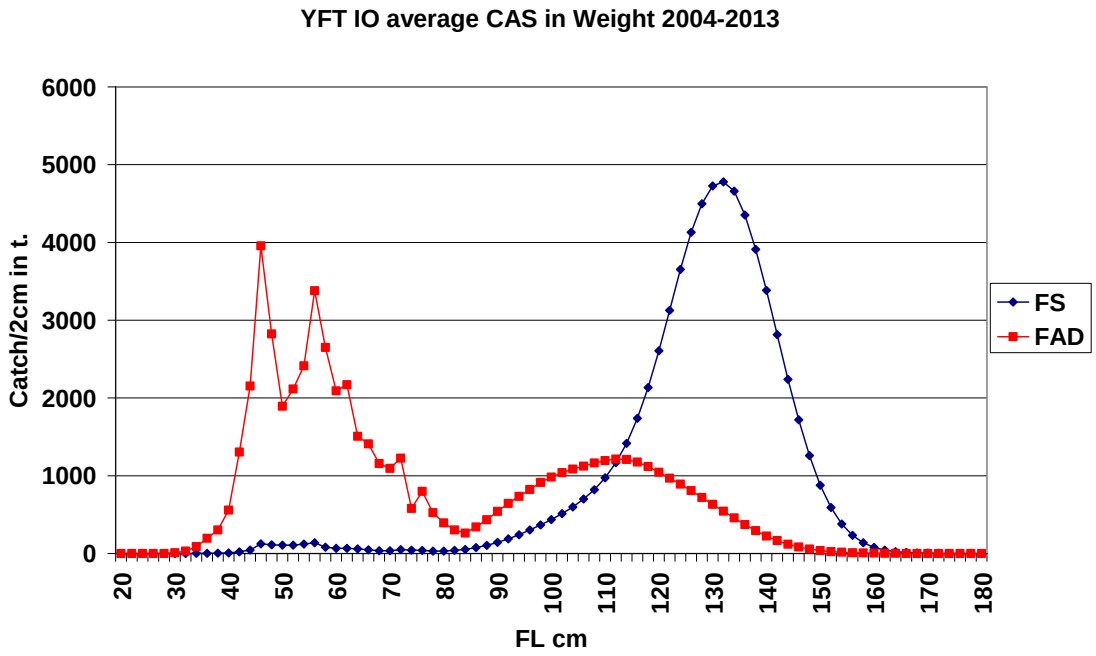


Figure 5 shows the predictions for the 2 sub-models (bottom) and their product (top). The 95% CI are also shown.

Annex I

Free schools YFT are most often caught at large sizes in a range between 1 m and 1.6 m (most of these tunas being potential spawners), while YFT caught in association to FADs are most often showing a bimodal structure: a 1st mode of small individuals (most often dominant in weight) between 40 and 80 cm and a 2nd mode of large individuals in a range between 80 and 140 cm. The following graph is based on IOTC catch at size data



Annex II: Residuals

Free School related Sets

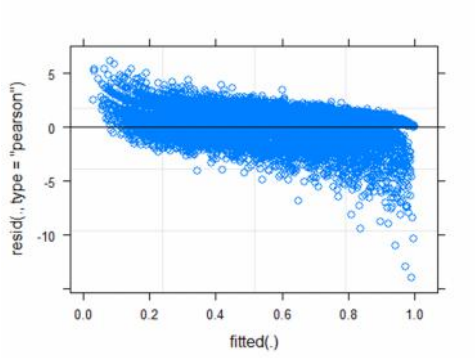


Figure 1 Plot of the residuals of the binomial model for the probability of a positive set.

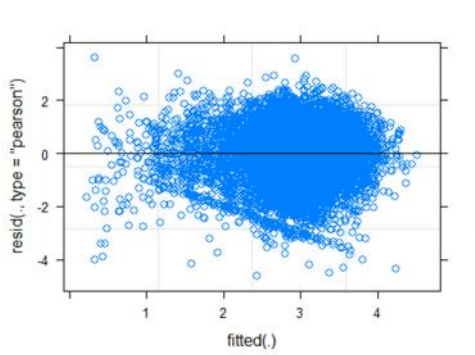


Figure 2 Plot of the residuals of the log-normal model for the catch per positive set.

Floating Object Related Sets

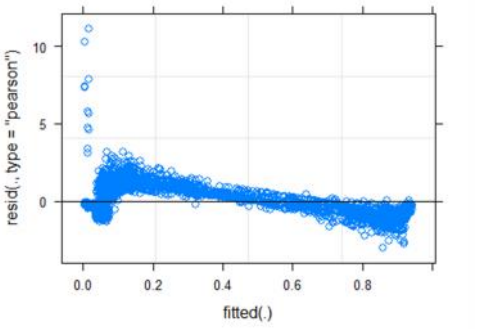


Figure 3 Plot of the residuals of the binomial model for the probability of a positive set.

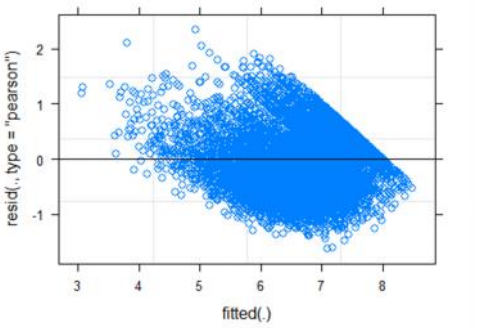


Figure 4 Plot of the residuals of the log-normal model for the catch per positive set.