TESTING A BYCATCH ESTIMATION TOOL USING SIMULATED BLUE MARLIN LONGLINE DATA

Elizabeth A. Babcock[1] and C. Phillip Goodyear[2]

SUMMARY

The species distribution model and longline simulator developed by Goodyear et al. (2017) were used to generate simulated longline sets for three fleets, catching both swordfish and blue marlin. These simulated data were used to test the effectiveness of a bycatch estimation tool in development. The tool allows for semi-automated model selection using information criteria to select the best set of predictor variables, and using cross validation to choose between Tweedie, negative binomial, delta-lognormal, and delta-gamma models. The simulated data allowed for a nuanced evaluation of the model decisions, such as whether to use trips or sets as a sample unit. As new functions are added to the bycatch estimation tool, the simulated data will continue to be key to adequate testing.

KEYWORDS

Bycatch, catch statistics, simulation, model testing

1. Introduction

The species distribution model (SDM) developed by Goodyear et al. (2017) generates a 3-dimensional distribution of blue marlin and swordfish throughout the Atlantic Ocean based on the habitat preferences of the species. Simulated longline sets are then generated by distributing hooks throughout the habitat of the species, consistent with the distribution, gear, hooks between floats, use of lightsticks and other characteristics of historical longline fishing fleets. These simulated data sets have been used to test algorithms for estimating indices of abundance based on catch per unit effort (CPUE) incorporating environmental and gear data (Forrestal et al., 2019). The simulated data was also used to evaluate different treatments of spatiotemporal variation, designs of observer programs, and potential bias caused by the possibility that the presence of observers might change fishing behavior (Güss et al., 2019). The complexity and realism of the simulated environment allows it to be used as an operating model to simulation test realistically complex data generation and analysis methodologies.

These simulated data were used to test the effectiveness of a bycatch estimation tool in development by Babcock (2021). The tool allows for semi-automated model selection using information criteria to select the best set of predictor variables and cross validation to choose between Tweedie, negative binomial, delta-lognormal, and delta-gamma models. The simulated data allows for evaluation of model decisions, such as whether to use trips or sets as a sample unit, how to group data in categorical variables, and whether to include environmental data.

2. Methods

The bycatch estimation code estimates total bycatch as follows. First, mean catch per unit effort (CPUE) of observed sample units (trips or sets) is estimated from a linear model with predictor variables in R (R Core Team, 2020). The observation error models used are delta-lognormal, delta-gamma, negative binomial (from either glm.nb in the MASS library or glmmTMB, nbinom1 and nbinom2) and Tweedie (from cpglm or glmmTMB) (Brooks et al., 2017; Dunn and Smyth, 2005; Venables and Ripley, 2002; Zhang, 2013). Within each observation error model group, potential predictor variables are chosen based on the user's choice of information criteria (AICc, AIC or BIC). The user specifies a most complex and simplest model, and all

[1] Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Cswy., Miami, FL 33149. USA. ebabcock@rsmas.miami.edu

[2] 686 Hickory LN, Havana, Florida 32333, USA. phil_goodyear@msn.com.

intermediate models are considered using the information criterion with the dredge function in the MuMIn library (Barton, 2020). The best candidate models in each observation error group are then compared using 10-fold cross-validation to see which observation error model best predicts CPUE. The best model according to cross validation is the one with the lowest root mean square error (RMSE) in the predicted CPUE and mean error (ME) closest to zero. Potential models that failed to converge, could not be fit due to insufficient data (e.g. no positive observations in some year prevents fitting the delta models), or produced results with unreasonably high CVs (>10) in the annual bycatch predictions are not used in cross-validation.

For the best model in each observation error model group, the total bycatch is estimated by predicting the catch in all logbook trips or sets (i.e., the whole fishery) from the fitted model and summing across all effort in each year. The catch in each sample unit is predicted directly by the negative binomial models. Tweedie models predict CPUE, which is then multiplied by effort. Delta-lognormal and delta-gamma models have separate components for the probability of a positive CPUE and the CPUE, which must be multiplied together (with appropriate bias corrections) and multiplied by effort to get the total catch. Catch is predicted in each trip (or set) in the logbook data using the model fitted to the observer data, and catch is summed across trips to get the total catch in each year. Because catch is being predicted in each trip, the variance of the prediction is calculated as the variance of the prediction interval, which is the standard error squared of the estimated catch plus the residual variance. Variances are summed across trips to get the total variance of the predictions in each year.

The software also calculates an annual abundance index from the same models that were selected for bycatch estimation. An annual abundance index is calculated by setting all variables other than year, and any variables required by the user to be included in the index (e.g. region or fleet) to a reference level, which is the mean for numerical variables or the most common value for categorical variables. The index and its standard error are then predicted at these reference levels.

Summary plots include the predicted number of positive trips plus and minus the standard deviation of the prediction, residual plots, residual qqnormal plots, and residuals calculated using the DHARMa R library (Hartig 2020). The DHARMa library uses simulation to generate scaled residuals based on the specified observation error model so that the results are more clearly interpretable than ordinary residuals. DHARMA draws random predicted values from the fitted model to generate an empirical predictive density for each data point and then calculates the fraction of the empirical density that is greater than the true data point. Values of 0.5 are expected, and values near 0 or 1 indicate a mismatch between the data and the model. Particularly for the binomial, negative binomial and Tweedie models, in which the ordinary residuals are not normally distributed, the DHARMa residuals are a better representation of whether the data are consistent with the assumed distribution.

For cross-validation, the observer data are randomly divided into 10 folds. Each fold is left out one at a time and the models are fit to the other 9 folds. The same procedure described above is used to find the best model within each observation error group using information criteria and the MuMIn library. The fitted model is used to predict the CPUE for the left-out fold, and the model with the lowest mean RMSE across the 10 folds is selected as the best model. Mean error is also calculated as an indicator of whether the model has any systematic bias.

As a test of the bycatch estimation methodology, we estimated the total bycatch of blue marlin in the simulated dataset. The simulated data are generated by set, while observers are allocated by trip. Thus, to generate correlated sets within simulated trips, we used the method of Grüss et al. (2019) to allocate sets to the same trip if they were in the same gear, month and spatial area (5 x 5 squares). Trips with more than 100 sets were randomly allocated to different trips so that the median trips had about 20 sets. This algorithm allowed for correlation among sets in the same trip to introduce potential clustering bias into the simulated observer data. We then simulated 5 percent observer coverage by trip. We used the bycatch tool to estimate the total bycatch and an index of abundance. Catch was in number of blue marlin caught per set, and effort was in 1000 hooks. We fitted the data aggregated at the level of trip, as might be necessary if effort data were reported by a logbook program that was not specific to sets. We also fit some models to the set by set data. For the set by set analysis, the variables considered in the model were year (1958-2018), fleet (3 levels), hooks between floats (hbf), area (North vs. South Atlantic), season (months 1-3,4-6,7-9 and 10-12), a year:area interaction in case trends were different in the two areas, and the habitat suitability of the location for blue marlin (w.BUM). The habitat suitability is a variable that is generated by the species distribution model based on temperature and other environmental variables. For the trip by trip analysis, we used median hbf and w.BUM for the input variables, since these would not be available at a set by set level. CPUE was the total blue marlin catch divided by the total number of hooks (/1000) in a trip or set. The goal of this comparison was validate the bycatch estimation tool, and to determine whether bycatch estimates could be improved by having set-by-set data rather than aggregating by trip in a hypothetical observer and logbook program for estimating total bycatch.

3. Results

The simulated dataset included three fleets, with data spanning from 1958 to 2018, where swordfish CPUE was very different between fleets, but blue marlin CPUE was similar across fleets (Figure 1). With 5% simulated observer coverage and data aggregated at the trip level, the fraction of observed trips that caught at least one blue marlin was quite high in the early years of the fishery, but decreased with time (Table 1). Raw CPUE, and the total bycatch estimated by a simple ratio estimator stratified only by year decreased over time as well.

For the trip-by-trip data, the cross validation found that the model that best predicted CPUE was the delta-gamma model (Table 2, Figure 2). The negative binomial models and the Tweedie also performed well. However, the delta-lognormal model had a much higher RMSE and ME. The information correctly disregarded the interaction between area and year, because the trend was the same over time in both regions in the simulated data. All the of the models, except the delta-lognormal, predicted very similar trends in both the total estimated catch of blue marlin (Figure 3) and index of abundance in both the North and South Atlantic (Figure 4). The DHARMa residuals showed adequate model specification in the binomial models and a slight tendency toward overdispersion in the gamma model (Figure 5). Comparing the predicted total blue marlin catch from the delta-gamma model to the actual total catch in the simulated data, the patterns are very similar (Figure 6a). However, some of the true data points are not within the 95% confidence intervals, implying that the bycatch estimation tool may be slightly underestimating variances. Comparing the true total catches to the delta-lognormal predicted total catches, shows much less overlap between the predictions and the true values, indicating that the cross-validation was correct to identify the delta-lognormal as an inferior predictive model (Figure 6b).

When a delta-gamma model was fitted to the set-by-set data, it gave similar results in the predicted total bycatch (Figure 7). However, the confidence intervals of the annual totals were even more narrow due to the increased sample size when using sets rather than trips as a sample unit.

Discussion

The simulated longline dataset provides a useful operating model for simulation testing bycatch estimation methodologies. For the trip-by-trip data, comparing the predicted total blue marlin catch to the true value confirmed that cross-validation aimed at finding the model that best predicted CPUE in the left-out portion of the observer data was effective at identifying which models would perform better or worse at predicting total catch. We were also able to identify that the methods used to predict the variance of the predictions may be underestimating variance, particular if sets rather than trips are used as the sample unit, an issue that requires further study.

Improvements in the bycatch estimation methodology that are currently in development, based on this exercise, include making more of the components run in parallel to improve processing speed with large datasets, and adding elements that would be useful for abundance index estimation, such as random effects, GAMS, and spatial correlation.

4. Acknowledgements

E. A. Babcock's work was supported by NOAA, via the Cooperative Institute for Marine and Atmospheric Science. C.P. Goodyear's work was supported by The Billfish Foundation.

5. References

Babcock, E. A., 2021. Bycatch estimator user guide. https://github.com/ebabcock/Bycatch-estimation

Barton, K. 2020. MuMIn: Multi-Model Inference. https://CRAN.R-project.org/package=MuMIn.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling. The R Journal 9 (2): 378–400.

Dunn, P. K., and G. K. Smyth. 2005. Series Evaluation of Tweedie Exponential Dispersion Models. Statistics and Computing 15: 267–80.

Forrestal, F. C., M. Schirripa, C. P. Goodyear, H. Arrizabalaga, E. A. Babcock, R. Coelho, W. Ingram, M. Lauretta, M. Ortiz, R. Sharma, and J. Walter. 2019. Testing robustness of CPUE standardization and inclusion of environmental variables with simulated longline catch datasets. Fisheries Research 210:1-13.

Goodyear, C. P., M. Schirripa, and F. Forrestal. 2017. Longline data simulation: a paradigm for improving CPUE standardization. Collect. Vol. Sci. Pap. ICCAT 74:379-390.

Grüss, A., J. F. Walter, E. A. Babcock, F. C. Forrestal, J. T. Thorson, M. V. Lauretta, and M. J. Schirripa. 2019. Evaluation of the impacts of different treatments of spatio-temporal variation in catch-per-unit-effort standardization models. Fisheries Research 213:75-93.

Hartig, F. 2020. DHARMa: Residual Diagnostics for Hierarchical (multi-Level / Mixed) Regression Models. https://CRAN.R-project.org/package=DHARMa.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth Edition. New York: Springer.

Zhang, Y. 2013. Likelihood-Based and Bayesian Methods for Tweedie Compound Poisson Linear Mixed Models.Statistics and Computing 23: 743-757.

Table 1. Summary of the simulated observer and logbook data for blue marlin, aggregated by trips with 5% observer coverage. Cat Est is the blue marlin bycatch estimated by a simple ratio estimator stratified only by year.

Year	Obs Cat	Obs Eff	Obs trips	CPUE	Pos	Pos Frac	Effort	trips	trips Obs Frac	Cat Est	Cat se
1958	744	496	16	1.46	16	1	4,800	168	0.0952	7,200	322
1959	490	472	19	1.04	19	1	9,190	349	0.0544	9,540	554
1960	354	540	30	0.693	28	0.933	12,400	497	0.0604	8,150	620
1961	424	918	33	0.614	29	0.879	16,000	666	0.0495	7,390	301
1962	1,430	1,640	67	0.792	61	0.91	33,000	1,280	0.0525	28,800	908
1963	1,480	1,720	62	0.854	60	0.968	33,000	1,240	0.05	28,300	765
1964	2,510	2,320	98	1.06	93	0.949	51,000	1,980	0.0494	55,000	1,510
1965	1,840	3,550	130	0.526	119	0.915	58,500	2,260	0.0576	30,300	614
1966	1,400	1,930	66	0.644	53	0.803	32,300	1,300	0.0507	23,300	566
1967	487	867	42	0.506	31	0.738	18,700	776	0.0541	10,500	562
1968	506	1,140	38	0.384	32	0.842	18,100	739	0.0514	8,080	279
1969	468	748	34	0.478	26	0.765	16,800	690	0.0493	10,500	558
1970	269	770	42	0.34	26	0.619	19,200	849	0.0495	6,700	525
1971	312	1,030	55	0.376	36	0.655	25,900	1,150	0.0478	7,830	571
1972	133	745	35	0.17	21	0.6	18,900	892	0.0392	3,380	410
1973	94	777	32	0.172	19	0.594	14,700	664	0.0482	1,780	276
1974	115	591	35	0.167	20	0.571	16,800	757	0.0462	3,280	518
1975	214	1,430	62	0.14	38	0.613	25,200	1,100	0.0565	3,780	362
1976	240	1,050	47	0.29	28	0.596	18,200	833	0.0564	4,160	316
1977	227	1,160	45	0.178	26	0.578	15,400	676	0.0666	3,000	216
1978	92	882	38	0.159	22	0.579	16,000	675	0.0563	1,670	297
1979	135	1,210	45	0.104	24	0.533	20,600	781	0.0576	2,290	278
1980	211	1,200	42	0.175	29	0.69	35,000	1,100	0.0382	6,160	447
1981	334	1,910	72	0.186	42	0.583	44,700	1,420	0.0507	7,790	482
1982	894	3,550	81	0.269	62	0.765	56,500	1,490	0.0543	14,200	275
1983	316	1,790	49	0.156	34	0.694	32,300	970	0.0505	5,690	260
1984	561	2,630	61	0.249	49	0.803	46,300	1,200	0.0507	9,900	263
1985	687	2,910	81	0.235	59	0.728	49,400	1,380	0.0586	11,700	322
1986	510	2,470	72	0.238	49	0.681	38,800	1,280	0.0561	8,000	283
1987	356	1,640	95	0.259	47	0.495	39,300	2,060	0.0461	8,560	714
1988	690	3,470	135	0.25	79	0.585	62,100	2,630	0.0514	12,300	511
1989	889	3,510	156	0.315	94	0.603	84,300	3,270	0.0478	21,400	812
1990	843	4,710	162	0.265	94	0.58	89,800	3,220	0.0503	16,100	563
1991	843	3,880	168	0.306	89	0.53	93,100	3,220	0.0521	20,200	818
1992	745	4,230	154	0.159	75	0.487	84,800	3,150	0.0489	14,900	594
1993	1,030	5,470	155	0.205	98	0.632	113,000	3,240	0.0479	21,300	540
1994	1,020	5,730	190	0.323	106	0.558	105,000	3,370	0.0563	18,700	577
1995	1,220	6,380	181	0.233	113	0.624	109,000	3,470	0.0521	20,900	485
1996	1,020	5,990	209	0.236	128	0.612	122,000	4,020	0.052	20,700	689

1997	1,300	5,380	168	0.253	97	0.577	103,000	3,540	0.0474	24,800	555
1998	1,100	5,300	171	0.233	106	0.62	113,000	3,660	0.0467	23,300	629
1999	679	4,520	176	0.19	104	0.591	97,000	3,390	0.0519	14,600	706
2000	928	4,670	164	0.208	92	0.561	104,000	3,280	0.05	20,600	670
2001	825	3,990	136	0.158	85	0.625	90,800	2,940	0.0463	18,800	616
2002	467	2,280	109	0.181	58	0.532	72,500	2,470	0.0441	14,900	921
2003	606	4,350	140	0.135	67	0.479	88,900	2,780	0.0504	12,400	547
2004	614	3,700	130	0.129	76	0.585	98,400	2,910	0.0446	16,400	718
2005	483	3,540	109	0.151	69	0.633	88,100	2,740	0.0399	12,000	578
2006	418	4,560	124	0.101	66	0.532	83,400	2,590	0.0479	7,650	433
2007	423	3,520	112	0.094	55	0.491	78,200	2,420	0.0464	9,390	533
2008	403	4,870	145	0.0678	74	0.51	86,300	2,490	0.0582	7,150	477
2009	271	3,780	128	0.0878	57	0.445	70,900	2,210	0.0578	5,090	510
2010	269	3,350	94	0.0728	46	0.489	72,300	2,160	0.0436	5,800	456
2011	223	3,580	108	0.0686	53	0.491	64,900	2,070	0.0521	4,050	431
2012	159	3,060	111	0.0468	47	0.423	67,800	2,210	0.0502	3,520	593
2013	140	2,320	99	0.054	37	0.374	58,200	1,990	0.0498	3,520	685
2014	123	2,120	79	0.0607	33	0.418	57,200	1,970	0.0401	3,320	617
2015	83	2,340	79	0.0594	24	0.304	46,700	1,680	0.0469	1,650	439
2016	100	2,580	62	0.0347	31	0.5	43,500	1,010	0.0612	1,680	275
2017	48	1,450	44	0.0445	19	0.432	46,400	1,040	0.0423	1,530	504
2018	96	2,600	56	0.0366	33	0.589	49,500	1,190	0.047	1,820	281

Table 2. For data aggregated by trip, best models for of each type according to the BIC, along with root mean square error (RMSE) and mean error (ME) from the cross validation, where CPUE was predicted by a delta method for the Lognormal and Gamma (multiplying positive trip CPUE times the probability of presence from the binomial model).

Model	Formula	RMSE	ME
Binomial	habBUM + hbf + season + 1 + area + fleet + Year	NA	NA
Lognormal	habBUM + hbf + 1 + area + fleet + Year	0.783	NA
Gamma	1 + area + fleet + Year	0.412	0.0778
TMBnbinom1	hbf + season + 1 + area + fleet + Year + offset(log(Effort))	0.445	0.0017
TMBnbinom2	hbf + season + 1 + area + fleet + Year + offset(log(Effort))	0.443	0.0117
TMBtweedie	hbf + 1 + area + fleet + Year	0.441	0.0078

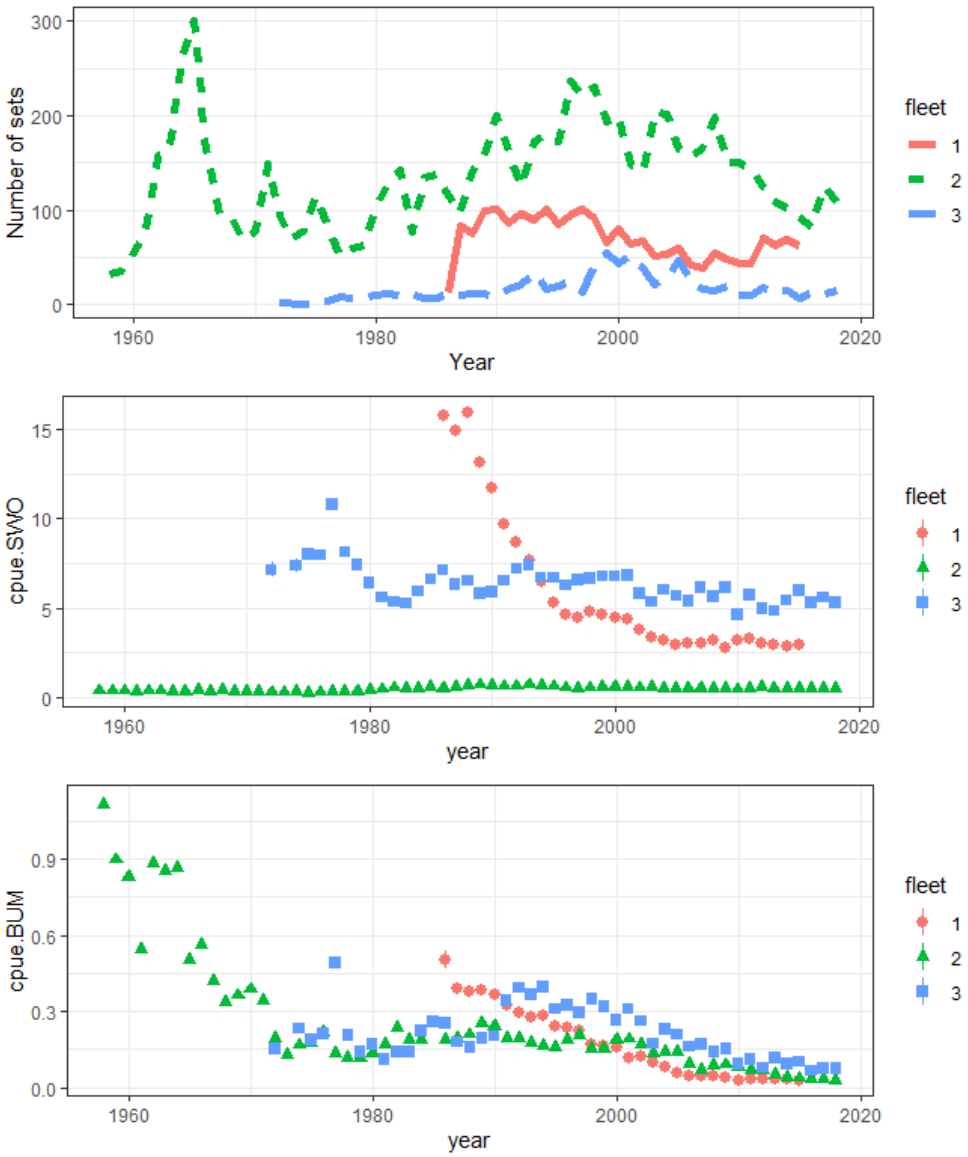


Figure 1. Total effort, and raw CPUE of swordfish and blue marlin in the simulated data set.

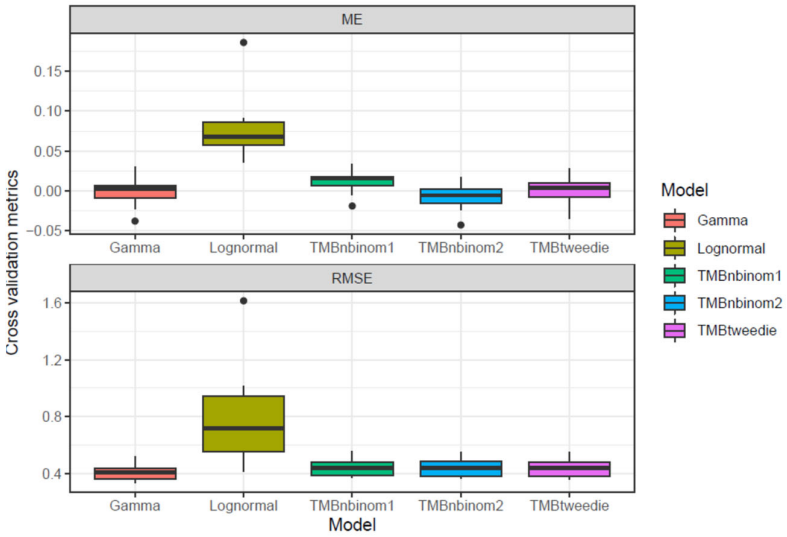
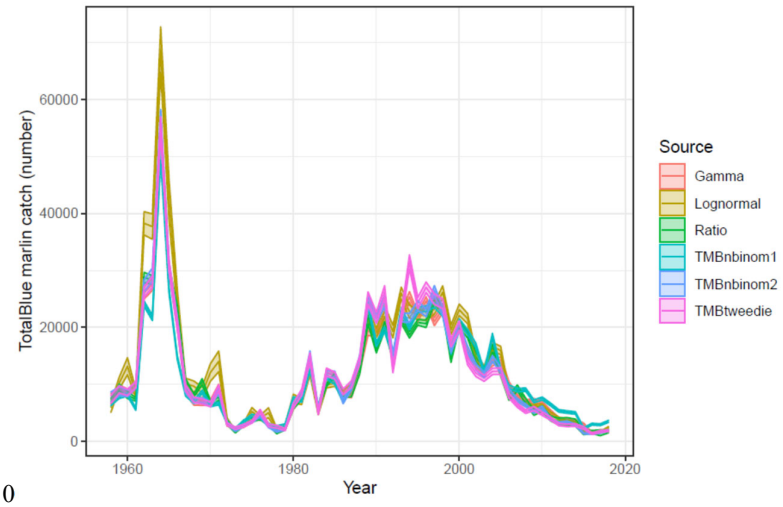


Figure 2. Cross validation results for each model group by trip.



0

Figure 3. Total bycatch estimate BIC best model in each error group.

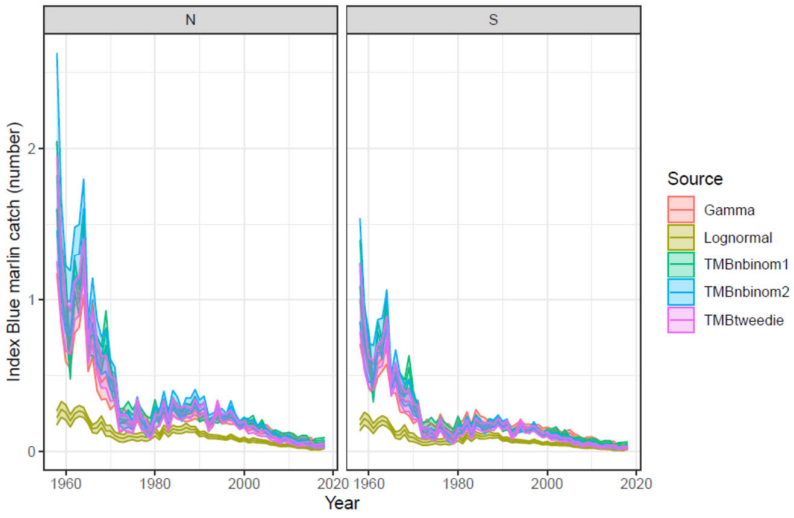


Figure 4. Estimated abundance indices for the selected BIC best model in each error group.

.

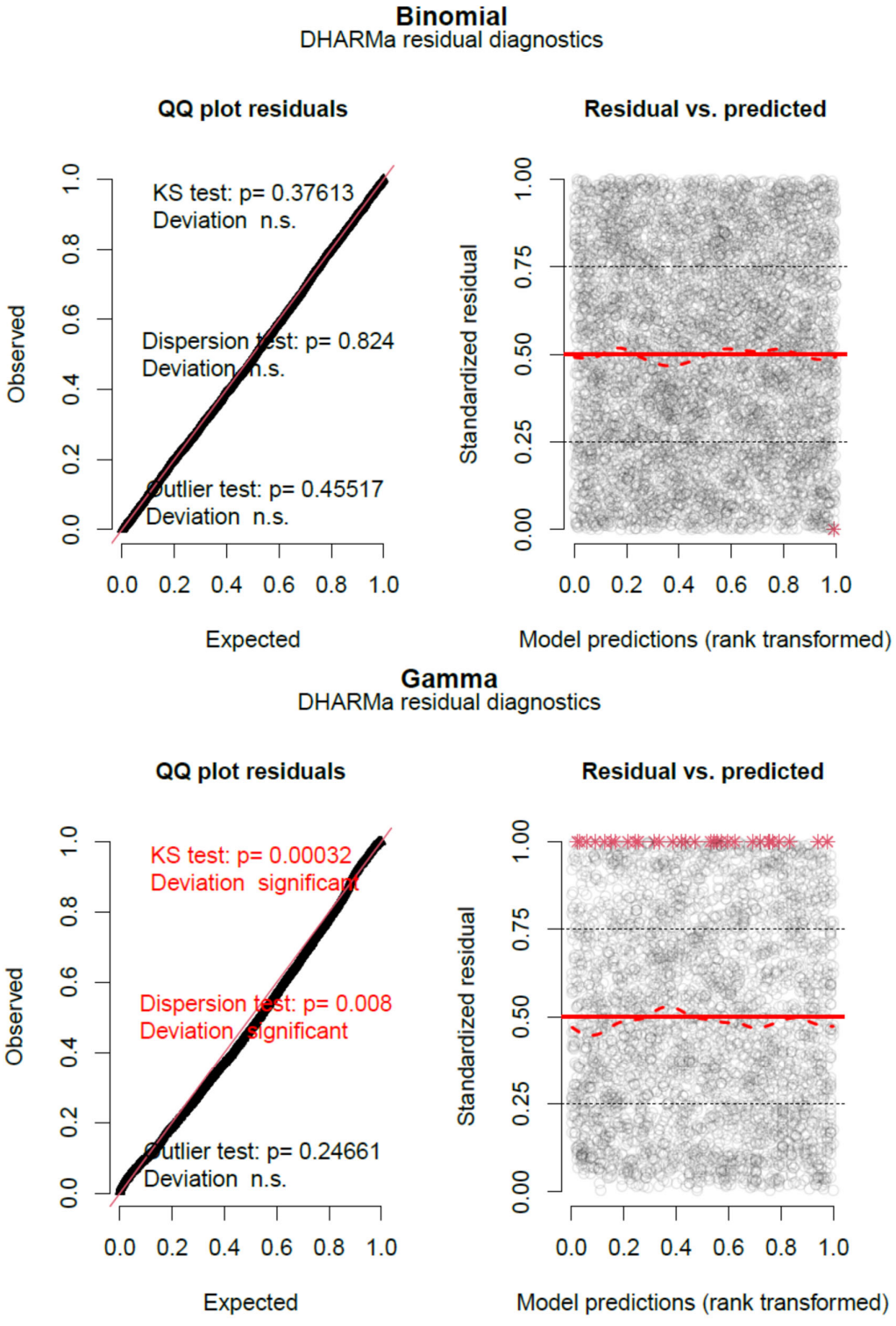
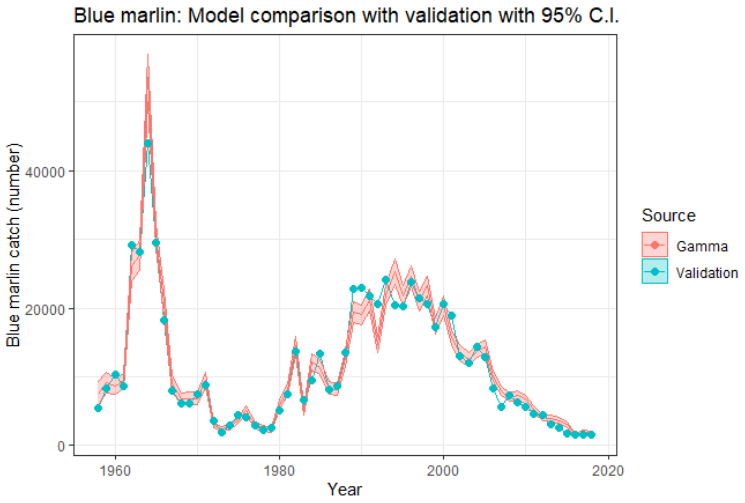


Figure 5. Diagnostics for best model by trip, which was a delta-model with both binomial and gamma components

(a) Best model by trip



(b) Worst model by trip

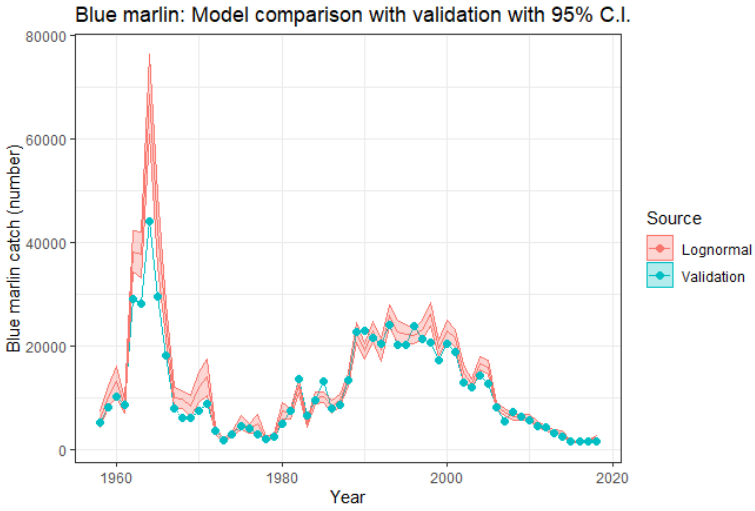


Figure 6. Comparison of model predicted total bycatch to actual total bycatch (labelled Validation) for (a) the model that performed best in cross-validation by trip, and (b) the model that performed worse in cross validation by trip.

0

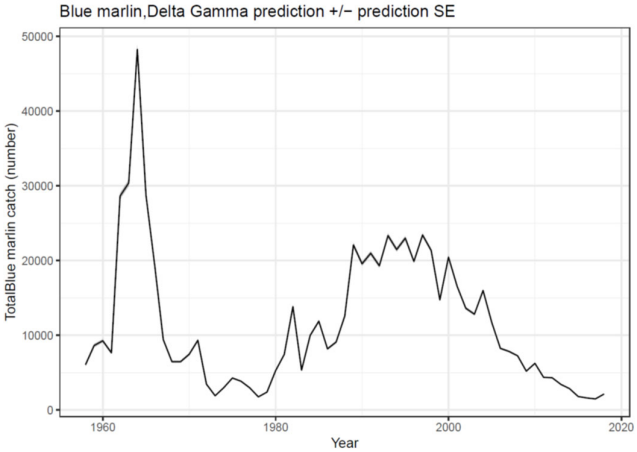


Figure 7. The best delta-gamma model prediction when data were entered by set rather than trip.