SCIENTIFIC COMMITTEE
NINTH REGULAR SESSION

Pohnpei, Federated States of Micronesia
6-14 August 2013

Swordfish CPUE for the Japanese charter fleet fishing off the west coast of the South Island of New Zealand

---

WCPFC-SC9-2013/SA-IP-08

Sam McKechnie, Simon Hoyle[1]

---

[1] Oceanic Fisheries Programme, Secretariat of the Pacific Community

1. Summary

This paper describes the analysis of swordfish CPUE from Japanese Charter Vessels fishing in New Zealand waters. Despite the small size of the fishery, this investigation was warranted due to the high quality of data for this fishery and the large swordfish that were consistently taken.

The dataset exhibits an extremely high proportion of zero-counts and low mean counts when positive sets occurred, which makes for challenging model fitting. This report presents standardizations based on Bayesian negative binomial, and zero-inflated negative binomial generalized linear models (GLMs).

The standardized index was similar to the nominal and varied substantially between years, including a five-fold increase over a three-year period (peaking in 1999/2000). Because the index relates to a time-series of relatively rare bycatch of swordfish in a fishery targeting southern bluefin tuna, we recommend using it only in sensitivity analyses within the assessment, and exercising caution when fitting models and interpreting results.

2. Introduction

The last stock assessment of swordfish in the southern WCPO was carried out in 2008 (Kolody *et al.* 2008), and will be undertaken again in 2013. New CPUE indices will be included in the upcoming assessment, and may include an index from the west coast of the South Island (WCSI), New Zealand. This paper presents the analysis of the catch rate of swordfish (*Xiphias gladius*) by Japanese charter vessels fishing for southern bluefin tuna (*Thunnus maccoyii*) within the New Zealand exclusive economic zone, with the aim of producing a time series of catch per unit effort (CPUE).

While this is a very small fleet of only 6 vessels with low swordfish catches, it is a relatively consistent fleet and the data are of high quality (100% operational logsheet data, very high observer coverage). The fleet also catches the largest swordfish taken in the southwest Pacific Ocean. A CPUE index that covers the largest fish is likely to be sensitive to changes in fishing mortality, and could therefore be valuable for the stock assessment.

Raw catch rates (nominal indices) may be affected by a range of factors over and above the abundance of the population. It is important to account for these effects by standardizing the CPUE time-series. This report presents attempts to standardize the operational level CPUE time series of the Japanese Charter fleet in New Zealand using generalized linear modelling techniques (and its variants) to provide a usable index of abundance for this region. The dataset exhibits an extremely high proportion of zero-counts and low mean counts when positive sets occurred, which makes for challenging model fitting. This report presents standardizations based on Bayesian negative binomial, and zero-inflated negative binomial generalized linear models (GLMs).

3. Methods

Operational data at the set level were available for the Japanese vessels in the NZ EEZ for the time period 1992-2012 (a total of 4,415 sets), with the exception of 1996 when no fishing occurred. Data were reported at a resolution of 1° spatial cell. Data were extremely sparse in the 1st and 4th quarters (Table 1) and were discarded, as were fishing activities outside the area of interest of the WCSI which covered most of the effort in the 3rd quarter. The resulting CPUE analyses will therefore be an annual, rather than quarterly, index.

The spatial extent of sets with positive catches was examined (Figure 1) with the aim of selecting a bounded region on the WCSI where positive catches (1 or more fish caught) occur. This was established by only including sets within the range of 40°S to 47°S, and to the west of 171°E. This left a dataset of 3,288 sets (2,832 zero catches and 456 positive catches) for which analyses were conducted.

A significant feature of the dataset is the high proportion of zero catches (0.86 over all sets; Figure 2, 3). Furthermore, when positive counts are observed they are generally very low (1-13 fish per set, mean=2.0, SD=1.7; Figure 4), which has consequences for the appropriateness of hurdle models such as the delta log-normal. The positive counts are not log-normally distributed (Figure 5) and nor is CPUE.

A zero-inflated negative binomial (ZINB) approach is often a suitable approach for datasets on non-target-species where encounter rates are low (e.g. Minami *et al.* 2007, Rice 2012, Smith *et al.* 2012). However, if the mean of positive counts is very low a simple (no zero inflation) negative binomial (NB) GLM may be adequate, as this distribution will predict a large number of zeros in these situations. The NB GLM can be represented by $y_i \sim \mathrm{NegBin}(\lambda_i, \delta)$, where y_i is the count for set *i*, λ_i is the mean and δ is the dispersion parameter. The mean of the negative binomial distribution is modelled using the usual linear predictor form

$$\log(\lambda_i) = \beta_1 + \beta_1 X_{1,i} \dots \beta_n X_{n,i}$$

where β are coefficients and X are explanatory variables.

If there are more zeros in the dataset than can be described by the simple NB GLM then further complexity must be added to the model. This is can be achieved using ZINB models. Firstly, the counts are modelled as negative binomial as a function of a set of explanatory variables using a GLM formulation. Then the occurrence of extra zeros (over and above the zeros expected to occur given the negative binomial distribution) is also modelled using a GLM formulation as a function of explanatory variables, which can be the same or different to those in the negative binomial section of the model.

The model is therefore based on

$$P(y_i|\theta) = \begin{cases} \pi_i + (1 - \pi_i) g(y_i| \theta) & y = 0, \\ (1 - \pi_i) g(y_i| \theta) & y = 1,2, \dots \end{cases}$$

where θ are model parameters, π_i is the probability of an excess zero and $g(y_i|\theta)$ is a distribution for the counts, which is assumed to be negative binomial here as there is likely to be overdispersion in the data that would not be accommodated by a more simple Poisson distribution. The model proceeds by modelling the zeros using the usual GLM formulation

$$\operatorname{logit}(\pi_i) = \alpha_0 + \alpha_1 W_{1,i} \dots \alpha_n W_{n,i}$$

where α are coefficients and W are explanatory variables. The mean of the NB count distribution (λ) is similarly modelled

$$\log(\lambda_i) = \beta_1 + \beta_1 X_{1,i} \dots \beta_n X_{n,i}$$

where β are coefficients and X are explanatory variables. Once the model has been fitted and parameters estimated, the standardized index of abundance can be calculated by multiplying the mean of the count distribution (λ) by one minus the probability of an excess zero.

Several explanatory variables that potentially influence swordfish catch rates were present in the dataset. This included the categorical variables; year (n=20), vessel (n=6), area (n=5; 5° cell where the set occurred; 5° cells were used due to high numbers of missing values for 1° cells in some years) and start time (n=12; the time that the set was made, binned into 12 two-hourly categories). The measure of effort used was the number of hooks per set which was added to the model as a continuous variable (an offset; log-transformed). In addition, moon phase was calculated based on the date of the set using the R function `moonphase` (from the package `r4mfcl`). This was transformed to be days from the full moon and was binned into four equal length categories.

A Bayesian approach was adopted using the software WinBUGS 1.4. This necessitated formulating prior distributions for all model parameters with attempts made to keep them uninformative by specifying dispersed normal distributions N(0,100) with means of zero and large variance, for all coefficients, and a Gamma(0.0001,0.0001) distribution for the negative binomial dispersion parameter (Martin *et al.* 2005, Smith *et al.* 2012). Selection among models with different explanatory variables proceeded using forward and backwards selection using DIC as a selection criterion, in addition to direct examination of the parameter estimates.

Model adequacy was assessed by comparing replicate datasets simulated under the model (from the posterior predictive distribution) with the observed dataset. If the replicated datasets (or functions of them e.g. test statistics) are similar to the observed dataset then the model is adequate. For example, if the data is from an approximately negative binomial distribution but a Poisson GLM is fitted then the replicated datasets will be underdispersed when plotted next to the observed dataset.

4. Results

Due to the very low mean of the positive counts observed, the simple negative binomial GLM appears to be an adequate model for this dataset. Datasets simulated under the model (from the posterior predictive distribution) were very similar to the observed dataset with respect to the full distribution (Figure 6), the distribution of positive counts (Figure 7) and the proportion of zeros in the dataset (Figure 8). This later feature is significant in that it appears that this dataset does not warrant adding a zero inflation component to the model. The best fitting model included all explanatory variables except for start time (year, log hooks, moon, area, vessel), however because most of these variables were relatively consistent over the time period the standardization produced an index that was very similar to the nominal index (Figure 9).

The CPUE index in the first three years was low but it increased quickly from 1998, reaching a peak in 1999 and 2000 at about five times the mean level (Figure 9). It then decreased again from 2001 and remained at a very low level between 2004 and 2012. In no years did the standardized index differ substantially from the nominal index and the inferred dynamics of abundance were consequently very similar.

The equivalent ZINB model was also fitted to the data for comparison purposes but it was not an improvement on the NB model (DIC=3,265 vs DIC=2,798 for the best NB model). While this model predicted replicate data similar to the observed data (Figure 10), the replicate data was not quite as consistent with the observed data, compared to the NB model, which can be seen graphically in a plot of the proportion of zeros and the total observed catch of swordfish (individual fish; Figure 11). Furthermore, the standardized index produced by this model is very similar to the NB model. The zero inflation component of this model has a constant probability of an excess zero and was simply

$$\operatorname{logit}(\pi) = \alpha_0$$

and models with any more complexity than this in the zero component (e.g. year effects, or linking the zero and NB components) produced unstable results (evidenced by lack of convergence and/or unrealistic parameter estimates.

5. Discussion

The analyses outlined produced a standardized index of swordfish abundance off the west coast of the South Island, New Zealand. The Bayesian NB model appeared to be a reasonable approach to modeling a challenging dataset, but the resulting standardization did not alter the index greatly and some of the worrying features evident in the nominal index therefore remain. Namely the rapid, substantial increase in catch rate of swordfish through 1998-2000 and the subsequent rapid decline to low catch rates from 2001 onwards. Given the large size and presumably old age of the swordfish taken in this fishery a rapid increase in abundance would require significant changes in either movement, recruitment, or fishing patterns. Swordfish catch rates are sensitive to operational factors such as hooks between floats, bait type and the use of light sticks, about which we lack information. The cause of the observed variation remains unclear and the fact that this dataset relates to the relatively rare bycatch of swordfish, and has very low sample sizes in many years for a CPUE analysis (especially for positive catches), suggest that the index should be treated with caution when inferring abundance from it.

Despite the extremely high proportion of zeros in the dataset ZINB models did not fit the data better than NB models due to the distribution of positive counts being restricted to very low numbers. The ZINB will be more suitable in situations where high proportions of zeros occur but the positive counts are higher than for west coast New Zealand swordfish. In the case of the NB model there are few substantial advantages of adopting a Bayesian approach, but for other models such as the delta log-normal the ease with which derived parameters such as a standardized index (and especially the associated estimates of uncertainty) can be calculated in a Bayesian framework may be very useful. For this reason the code for fitting the ZINB and NB model (which could be easily adapted to other situations) is provided in the appendix.

6. References

Kolody. D., Campbell, R., and Davies, N. (2008). A MULTIFAN-CL assessment of south-west Pacific swordfish 1952-2007. WCPFC-SC4-2008/SA-WP-6.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecological Letters 8: 1235-1246.

Minami, M., Lennert-Cody, C. E. (2007). Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. Fisheries Research 84: 210-221.

Rice, J. (2011). Catch per unit effort of silky sharks in the Western and Central Pacific Ocean. WCPFC-SC8-2012/SA-IP-11.

Smith, A. N. H., Anderson, M. J., Millar, R. B. (2012). Incorporating the intraspecific occupancy-abundance relationship into zero-inflated models. Ecology 93: 2526-2532.

7. Tables

Table 1: Total number of sets in NZ EEZ for each quarter/year. Note that no fishing occurred in 1996.

	Quarter			
Year	1	2	3	4
1992	0	154	20	0
1993	8	78	24	0
1994	7	96	0	0
1995	0	136	0	0
1996	0	0	0	0
1997	21	199	23	0
1998	0	135	22	0
1999	0	262	12	0
2000	0	185	14	0
2001	0	198	0	0
2002	5	229	0	0
2003	0	391	138	0
2004	1	315	34	0
2005	0	156	42	0
2006	0	150	48	13
2007	36	317	90	0
2008	0	167	0	0
2009	0	209	22	0
2010	0	144	0	0
2011	0	151	0	0
2012	0	163	0	0

Table 2: Number of sets/ year for each vessel for the final dataset for WCSI in quarter 2.

Vessel	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12
Ves 1	0	0	38	0	0	31	41	52	46	42	59	67	81	81	70	60	41	51	35	36	37
Ves 2	59	0	0	0	0	0	0	57	50	46	39	58	73	70	71	81	38	45	36	36	40
Ves 3	74	59	1	41	0	39	49	62	0	40	54	67	67	0	0	66	43	0	0	0	0
Ves 4	0	0	25	48	0	38	21	53	41	47	62	69	76	0	0	0	0	0	0	0	0
Ves 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	45	55	39	42	44
Ves 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	34	35	40

Table 3: Nominal and standardized indices of abundance of WCSI swordfish.

	Nominal Estimate	Standardized Estimate	SE
1992	0	0.016	0.017
1993	0.910	0.689	0.242
1994	0	0.005	0.018
1995	0.209	0.220	0.106
1996	NA	NA	NA
1997	0.846	0.760	0.194
1998	2.567	2.117	0.385
1999	4.645	5.123	0.525
2000	4.982	4.929	0.623
2001	1.792	2.144	0.334
2002	0.800	0.940	0.175
2003	0.561	0.573	0.108
2004	0.148	0.165	0.049
2005	0.354	0.392	0.116
2006	0.191	0.251	0.097
2007	0.347	0.253	0.058
2008	0.382	0.368	0.103
2009	0.097	0.084	0.038
2010	0.143	0.151	0.070
2011	0.207	0.247	0.100
2012	0.819	0.577	0.152

8. Figures

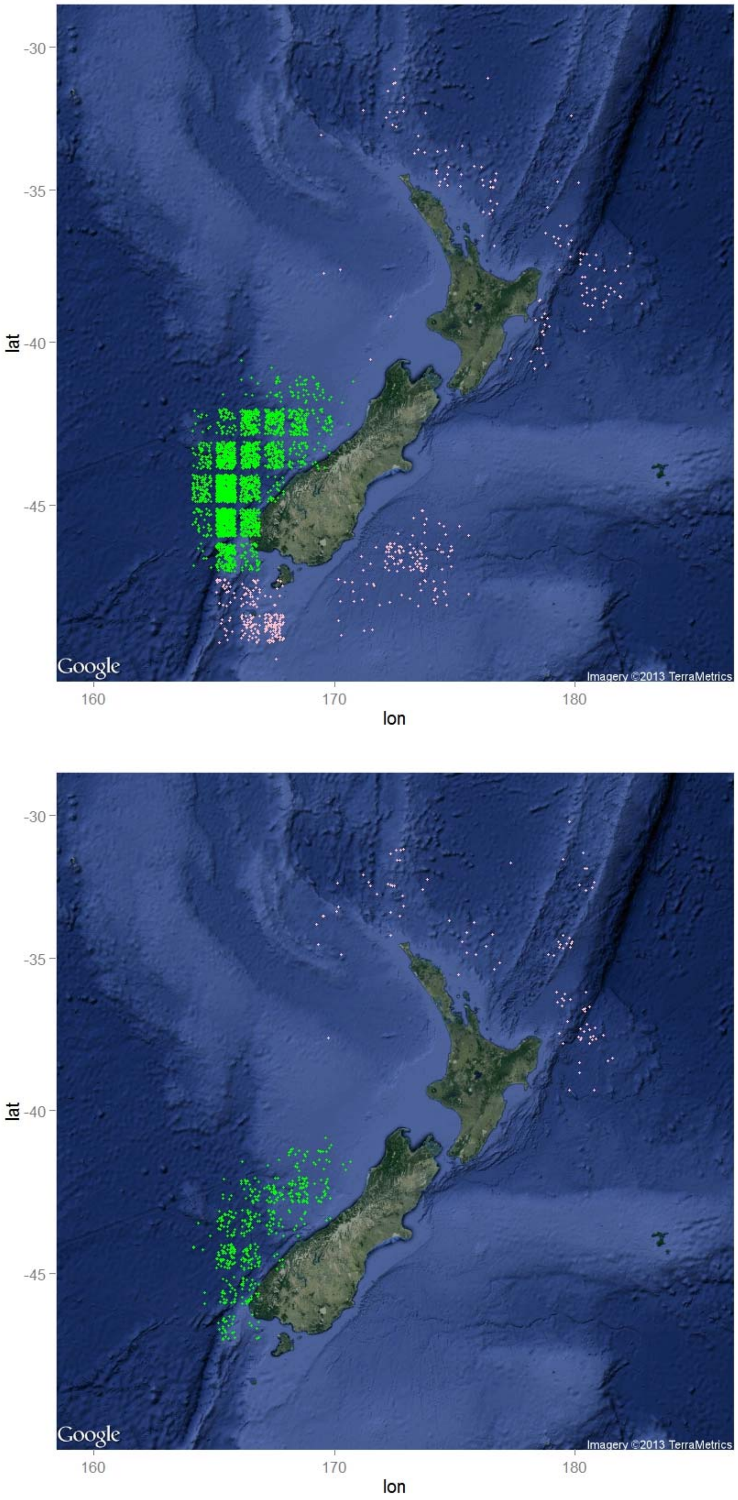

Figure 1: Distribution of sets in the New Zealand region in quarter 2. Each dot represents a set, the green dots are sets in the defined fishery region and the pink dots are the sets excluded from analyses. The top plot shows sets with zero catches and the bottom plot is the sets with positive catches. Data is at the 1 degree square scale but have been jittered to show the frequency within each cell.

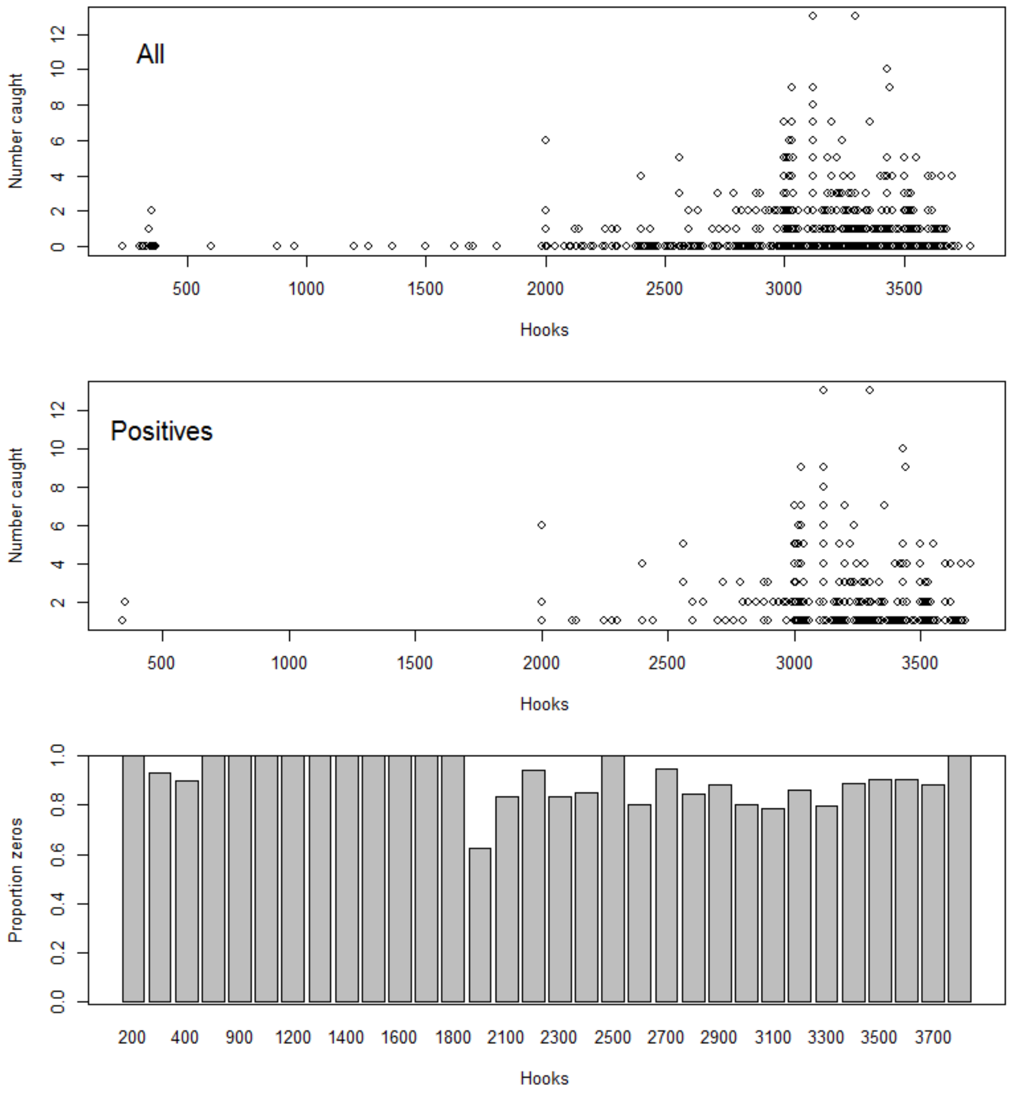

Figure 2: Relationship between catches and effort (hooks per set). Figures are for all sets (top), only sets where positive catches occurred (middle) and the proportion of sets that were zero catches (bottom; catches were binned into even width categories). Each dot represents one set.

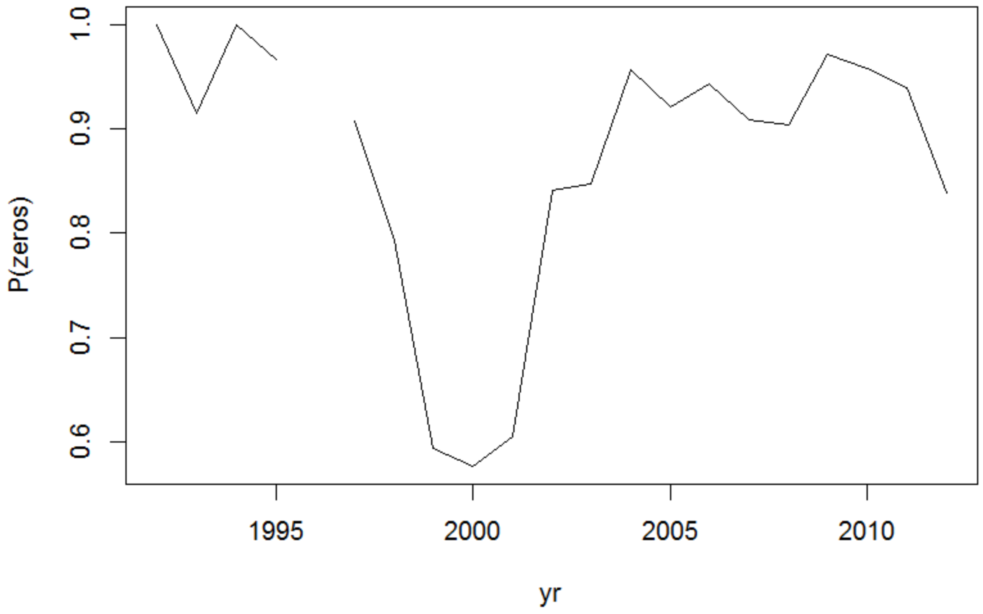

Figure 3: The observed proportion of sets that were zero catches for each year.

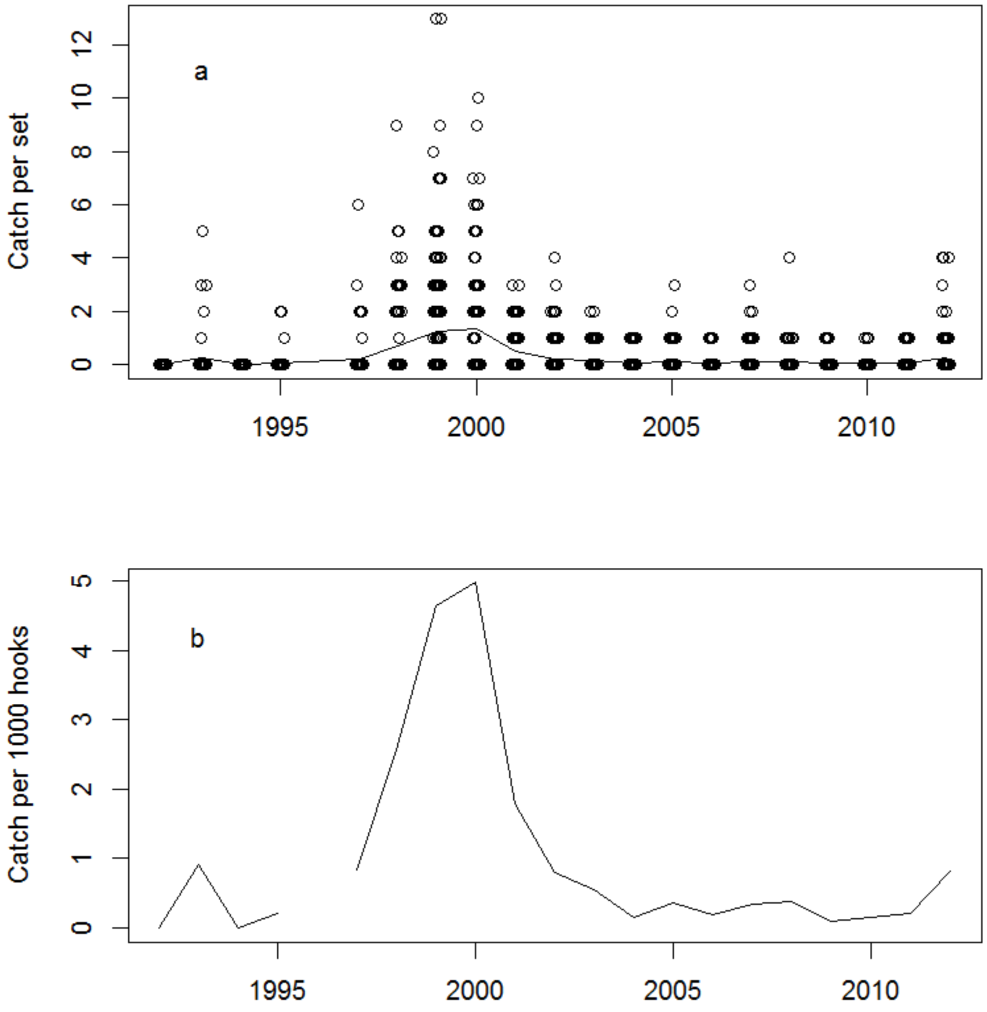

Figure 4: Raw annual catch rates. The top panel shows the catch per set for all sets in the final dataset (each dot represents a single set – jittered) with the line the yearly mean. The bottom panel is the nominal CPUE index (catch per thousand hooks).

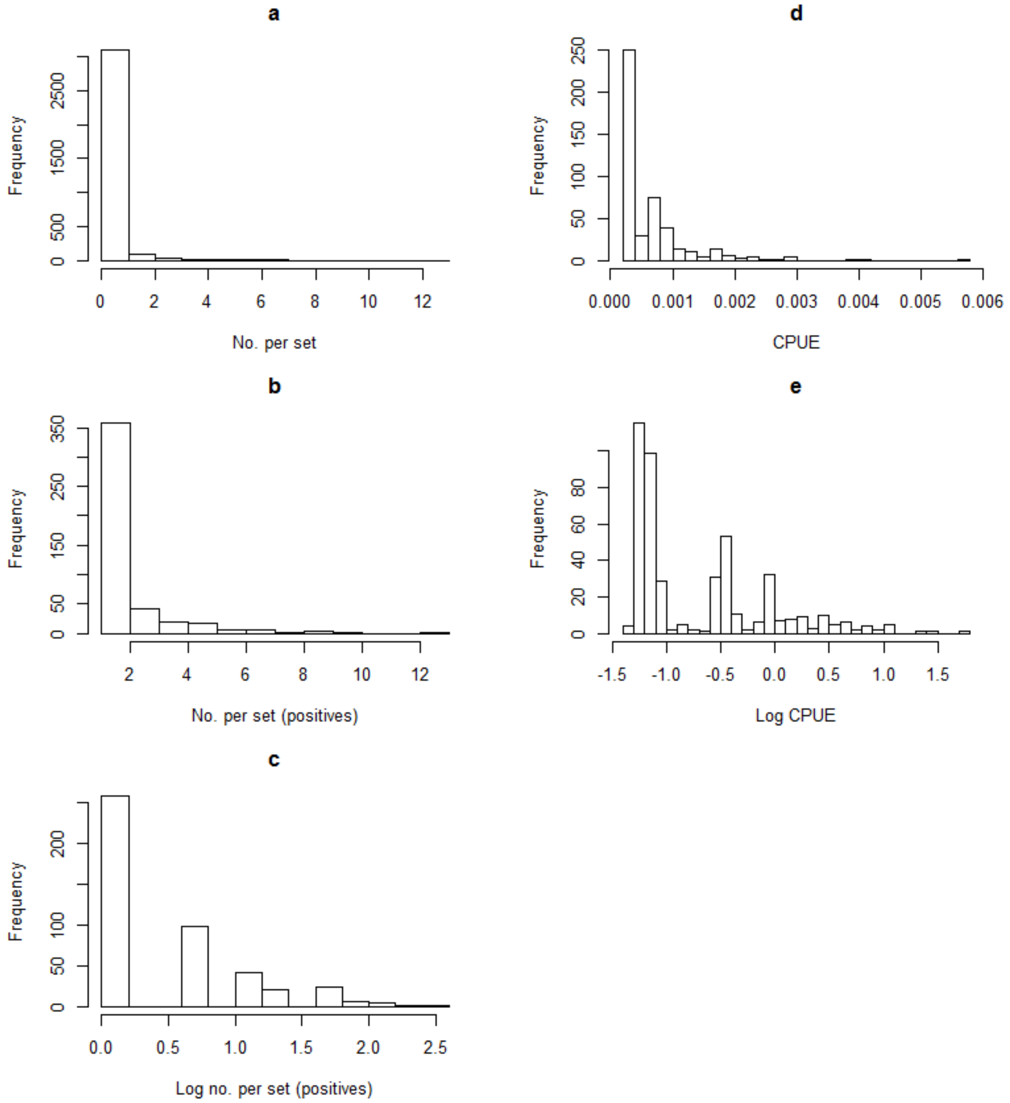

Figure 5: Histograms showing the distributions of catches per set (a-c) and CPUE (catches per thousand hooks; d-e). Note that neither the log(catches per set) or log(catches per hook) appear to be well described by a normal distribution.

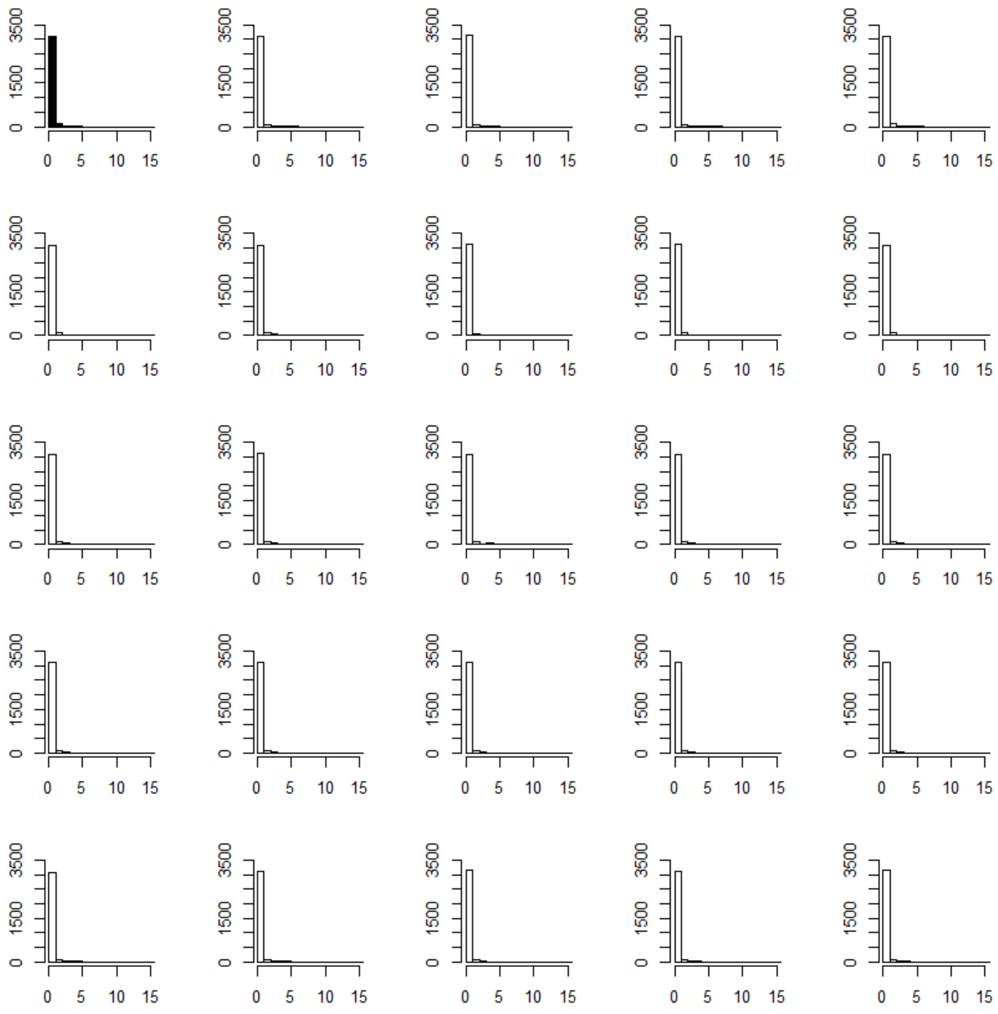

Figure 6: Distribution of observed counts (black histogram) and examples of counts for datasets simulated under the best NB model (from the posterior predictive distribution). This shows that the model appears to be adequate with respect to this test "statistic" as the observed and simulated distributions are very similar.

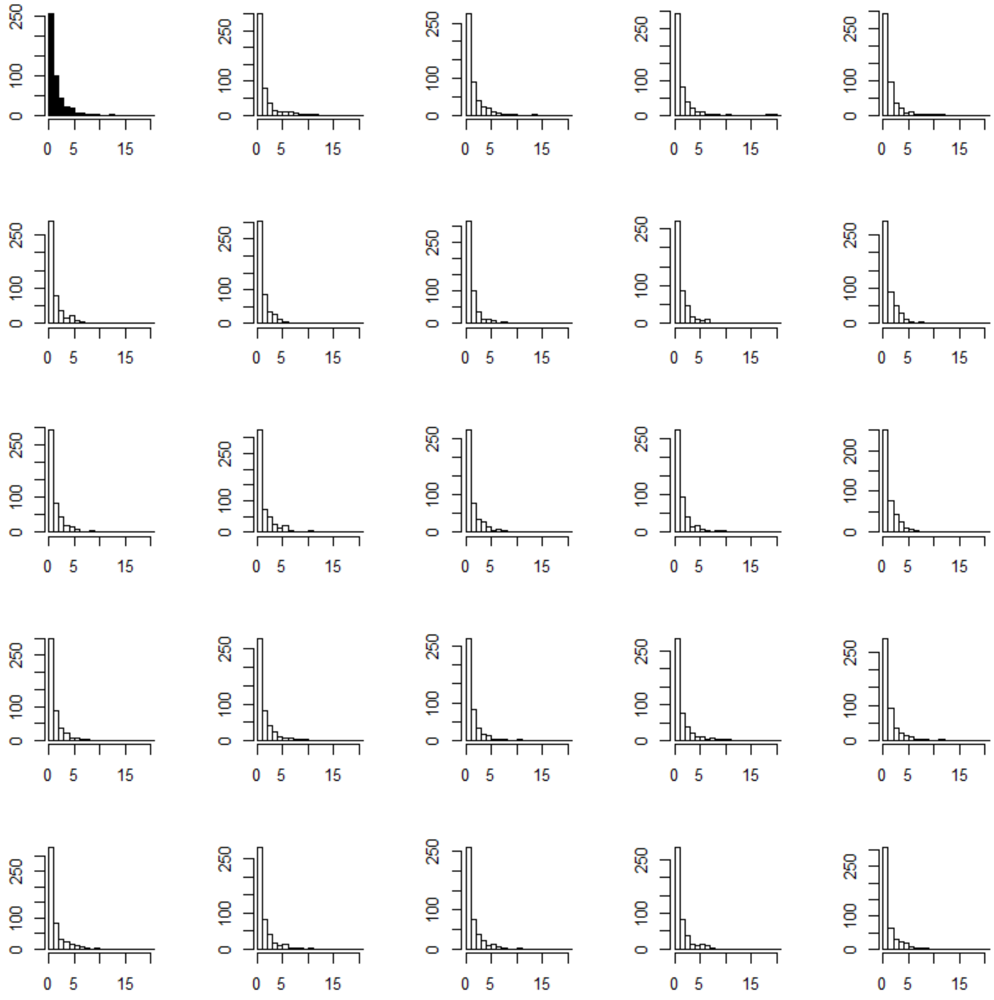

Figure 7: Distribution of observed positive counts (black histogram) and examples of positive counts for datasets simulated under the best NB model (from the posterior predictive distribution). The observed and simulated distributions appear to be very similar.

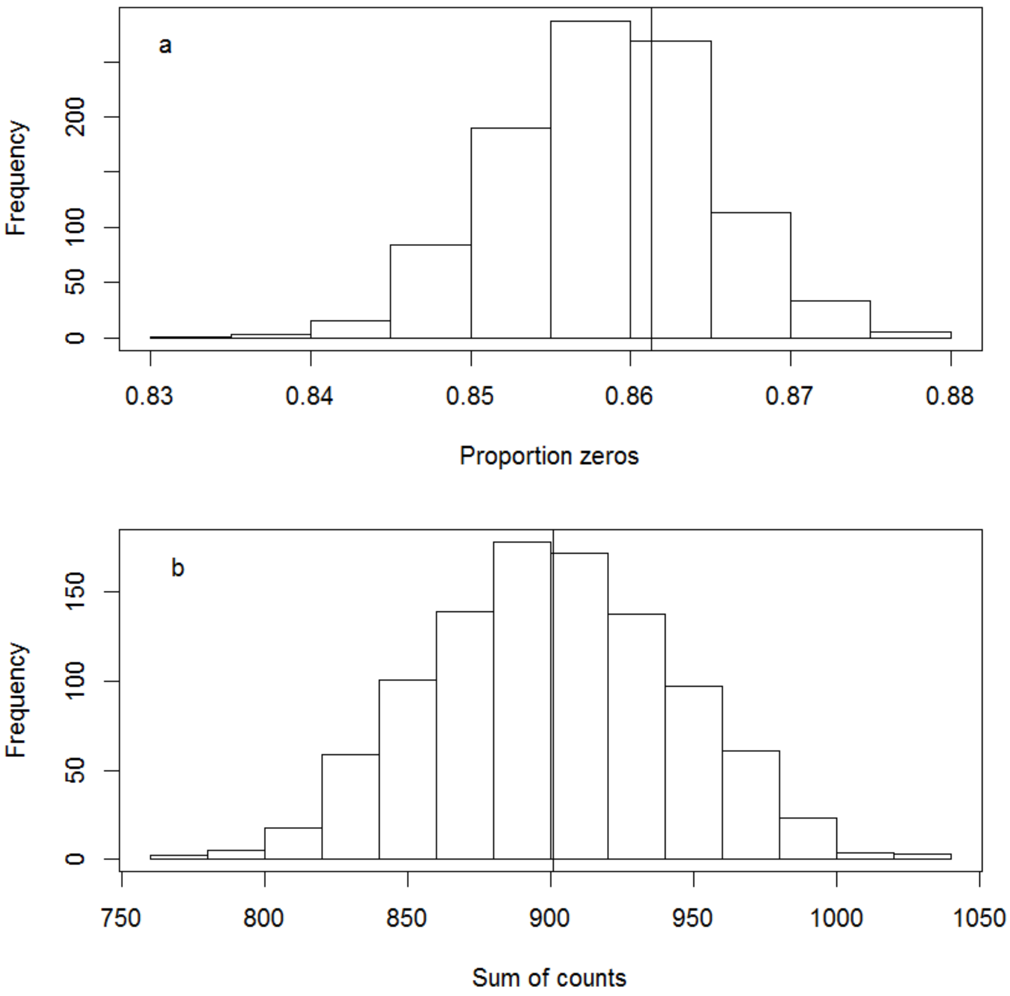

Figure 8: The observed proportion of zeros in the dataset (panel a – vertical line) and the observed total catch of individual swordfish over the dataset (panel b – vertical line) and the distributions of the same statistics for 1000 replicate datasets simulated under the best NB model. Because the replicated "test statistics" are not consistently higher or lower than the observed statistics, the model appears adequate in describing these aspects of the data.

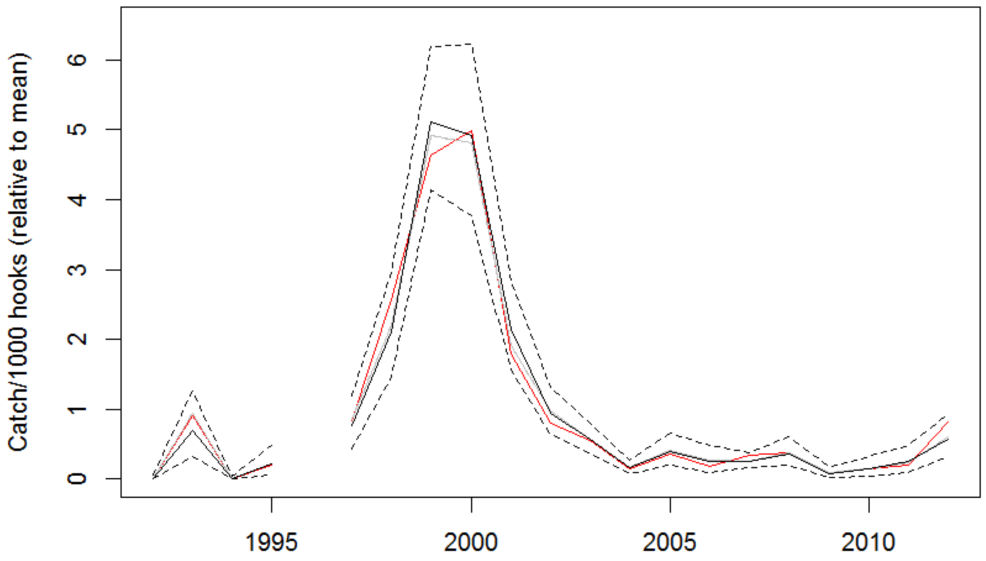

Figure 9: The nominal CPUE index (red line) overlaid with the mean standardized CPUE index for the best NB model (black line). Dotted lines are the limits of 95% credible intervals. The grey line shows the mean of the standardized index produced by the best ZINB model for comparison purposes.

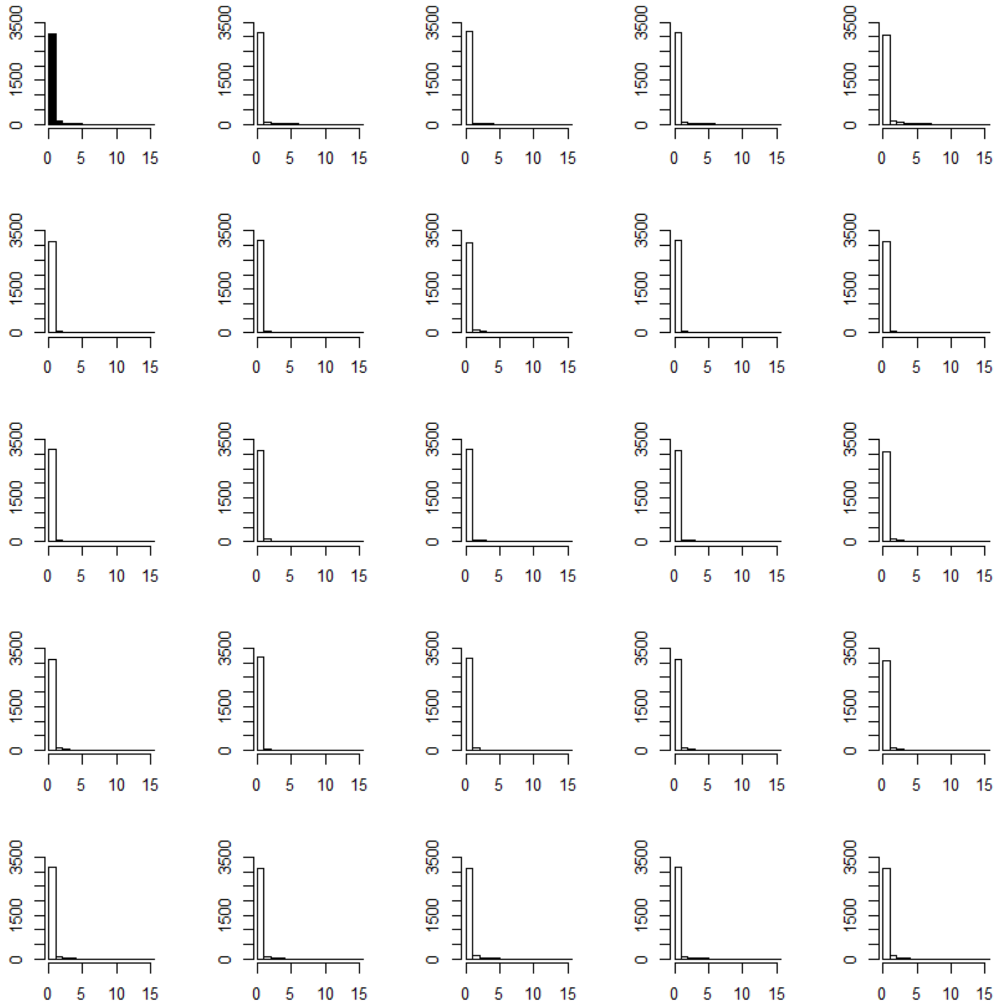

Figure 10: Distribution of observed counts (black histogram) and examples of counts for datasets simulated under the best ZINB model (from the posterior predictive distribution).

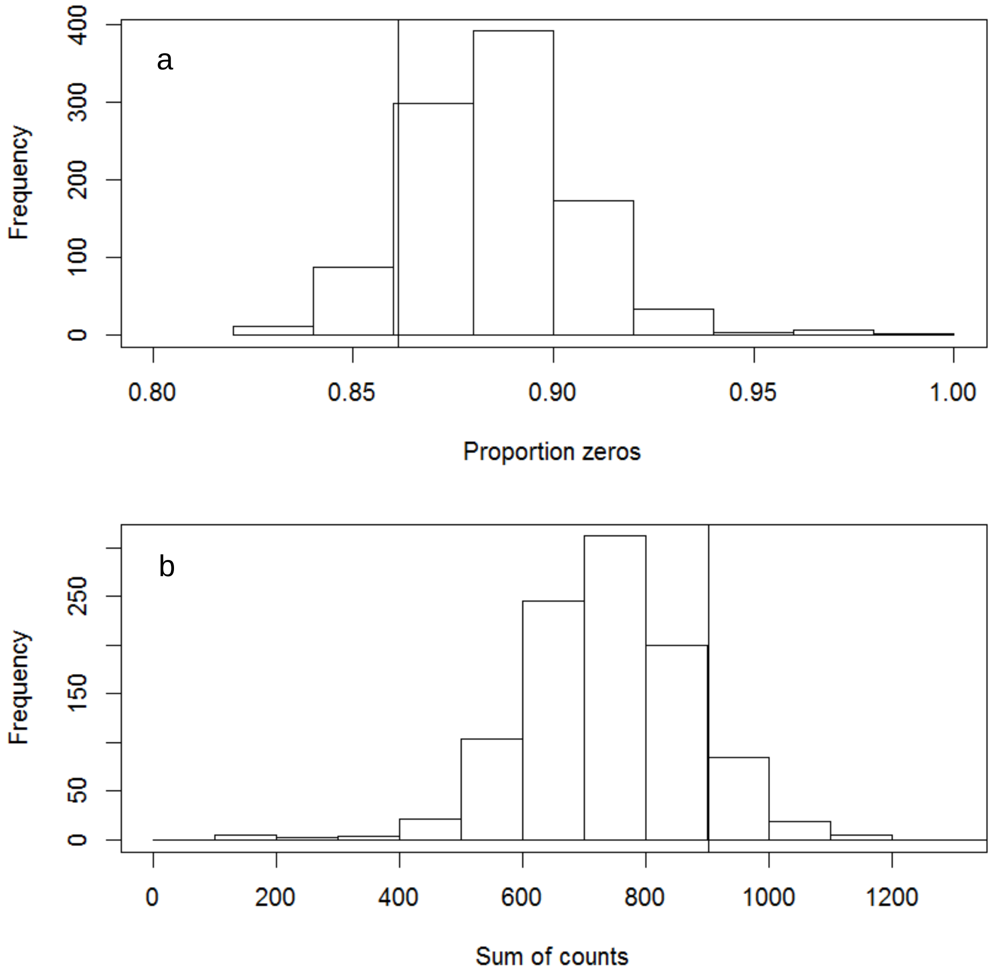

Figure 11: The observed proportion of zeros (panel a – vertical line) and the total catch of individual swordfish over the dataset (panel b – vertical line) and the distributions of the same statistics for 1000 replicate datasets simulated under the best ZINB model. Because nearly all the replicated "test statistics" are higher than the observed statistics, the model appears to be inadequate in describing the distribution of positive counts.

9. WinBUGS code for a NB model and a ZINB with an intercept-only excess zero component

9.1. Zero inflated negative binomial GLM with intercept-only excess zero component

```
model {
  for(i in 1:3288) {

#  Need to use the "zeros trick" to set up the negative log likelihood
       zero[i] <- 0
       zero[i] ~ dpois(FullNLL[i])

# The negative log likelihood for the zeros is the prob of excess zeros (p) + 1-p * prob of neg binom zeros
# D is the dispersion parameter of the negative binomial, L is the mean (usually denoted λ)
       NLLZero[i] <- -log(p[i] + (1 - p[i])*pow(1 + L[i] / D, -D))

# The negative log likelihood for non-zeros is 1- prob of excess zeros + neg binom prob function
# COUNT is the observed count of fish
       NLLNonZero[i] <- -(log(1 - p[i]) + (loggam(D + COUNT[i]) - loggam(D) - loggam(COUNT[i] + 1) +
                          COUNT[i] * log(L[i]) - COUNT[i] * log(D) - (D + COUNT[i]) * log(1 + L[i] / D)))

# Neg log likelihood depends on whether observed count is a 0 (Indicator=0) or 1 (Indicator=1)
       FullNLL[i] <- (1-INDICATOR[i]) * NLLZero[i] + INDICATOR[i] * NLLNonZero[i]

# Set up the linear predictor for the excess zeros – in this example only an intercept
       logit(p[i]) <- A0    # Note that this is just example code and throughout p does not need to be indexed by i but
if more complicated models are to be fitted then must be p[i], for example Smith et al. 2012 suggest linking the
zero inflation to the positives functionally - logit(p[i]) <- A0 + A1*LinPred[i]

# Set up the linear predictor for the mean of the negative binomial component – note log(hooks) is
# centred as speeds convergence
       log(L[i]) <- LinPred[i]
       LinPred[i] <- B0 + B1[YRS[i]] + B2*(logHOOKS[i]-mean(logHOOKS[])) + B3[MOON[i]] + B4[AREA5[i]]
                    +  B5[VES[i]] + B6[START[i]]

     logHOOKS[i] <- log(HOOKS[i])

     junk[i] <- AREA[i]     # Must use all data in winbugs, not a problem in openbugs though... so if some
                            # data are not used just make up a junk variable
  }

# For factors set the first level to zero to make model identifiable – can also be set to sum to zero to
# speed convergence in some cases e.g. for B1 it would be          B1[1] <- -sum(B1[2:20])
     B1[1] <- 0
     B3[1] <- 0
     B4[1] <- 0
     B5[1] <- 0
     B6[1] <- 0

# Priors for all the factors below
  for(i in 2:20) {
     B1[i] ~ dnorm(0.0, 0.01)
  }
```

```
   for(i in 2:4) {
       B3[i] ~ dnorm(0.0, 0.01)
   }

   for(i in 2:5) {
       B4[i] ~ dnorm(0.0, 0.01)
   }

   for(i in 2:6) {
       B5[i] ~ dnorm(0.0, 0.01)
   }

   for(i in 2:11) {
       B6[i] ~ dnorm(0.0, 0.01)
   }

  # priors for the other parameters here
  D ~ dgamma(0.0001,0.0001)
  A0 ~ dnorm(0.0, 0.01)
  B0 ~ dnorm(0.0, 0.01)
  B2 ~ dnorm(0.0, 0.01)

# Derived parameters here – calculate the standardized index
# Prob of an excess zero – if a more complex linear predictor is used this would be within the for loop
# below
       P.yr <- 1/(1 + exp(-(A0)))

   for(i in 1:20) {
# Predict the mean of the negative binomial component
       L.yr[i] <- exp(B0 + B1[i] + B2*(log(1500)-mean(logHOOKS[])) + B3[2] + B4[2] + B5[2] + B6[2])
# Predict the overall mean index
       mean.yr[i] <- (1 - P.yr) * L.yr[i]
# Standardise mean by the overall mean
       std.mean.yr[i] <- mean.yr[i] / mean(mean.yr[])

# Code below repeats with different levels of the covariates/factors in the prediction etc. to demonstrate that for
these models it produces the same index
       P.yr1[i] <- 1/(1 + exp(-(A0)))
       L.yr1[i] <- exp(B0 + B1[i] + B2*(log(3500)-mean(logHOOKS[])) + B3[4] + B4[4] + B5[4] + B6[4])
       mean.yr1[i] <- (1 - P.yr1[i]) * L.yr1[i]
       std.mean.yr1[i] <- mean.yr1[i] / mean(mean.yr1[])
   }

}


# Example of a row of data – area was added at two scales with AREA not used in the model above
# hence the junk variable was made in the model
YRS[] VES[] AREA[] AREA5[] HOOKS[] START[] MOON[] COUNT[] INDICATOR[]
3 1 29 5 2826 11 2 0 0

# Example of initial values – these are zeros for demonstration purposes but better to come up with
# reasonable values for specific problems – use different values for different chains.
list(D = 0, A0 = 0.1, B0 = -3, B1 =c(NA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), B2=0, B3=c(NA,0,0,0), B4=c(NA,0,0,0,0),
B5=c(NA,0,0,0,0,0), B6=c(NA,0,0,0,0,0,0,0,0,0,0))
```

9.2. Negative binomial GLM

```
# Most aspects of the more simple NB glm follow the ZINB above
model {
  for(i in 1:3288) {
      COUNT[i] ~ dpois(L[i])    # To model the NB parameterized by it's mean in WinBUGS requires using a mixture of the poisson and a gamma
      L[i] <- rho[i]*exp(LinPred[i])  # Linear predictor and extra gamma variation (equivalent to NB)
      LinPred[i] <- B0 + B1[YRS[i]] + B2*(logHOOKS[i]-mean(logHOOKS[])) + B3[MOON[i]] + B4[AREA5[i]] + B5[VES[i]]
      rho[i] ~ dgamma(A0,A0)  # Extra gamma
      logHOOKS[i] <- log(HOOKS[i])

      junk1[i] <- INDICATOR[i]  # Some of dataset unused so add  meaningless variables if using winbugs
      junk2[i] <- AREA[i]
      junk3[i] <- START[i]
  }

       B1[1] <- 0
       B3[1] <- 0
       B4[1] <- 0
       B5[1] <- 0

  for(i in 2:20) {
       B1[i] ~ dnorm(0.0, 0.01)
  }

  for(i in 2:4) {
       B3[i] ~ dnorm(0.0, 0.01)
  }

  for(i in 2:5) {
       B4[i] ~ dnorm(0.0, 0.01)
  }

  for(i in 2:6) {
       B5[i] ~ dnorm(0.0, 0.01)
  }


 # priors
 A0 <- exp(logA0)
 logA0 ~ dnorm(0,0.0001)
 B0 ~ dnorm(0.0, 0.01)
 B2 ~ dnorm(0.0, 0.01)

 for(i in 1:20) {
       L.yr[i] <- exp(B0 + B1[i] + B2*(log(1500)-mean(logHOOKS[])) + B3[2] + B4[2] + B5[2])
       std.mean.yr[i] <- L.yr[i] / mean(L.yr[])

       L.yr1[i] <- exp(B0 + B1[i] + B2*(log(3500)-mean(logHOOKS[])) + B3[4] + B4[4] + B5[4])
       std.mean.yr1[i] <- L.yr1[i] / mean(L.yr1[])
 }

}
```