A further update of the hake species splitting model.

by

OLRAC-SPS
Silvermine House
Steenberg Office Park
Tokai 7945

December 2012

In fulfilment of the 2nd client action plan for maintaining MSc certification

Summary

A revision of the species splitting formula for hake using updated survey data is presented. The equation reported here is based on a Generalised Linear Model with a binomial distribution function and a logit link function. The equation that is produced is the currently used logistic equation as a function of depth. Model selection is based on the predictive power of the model for a 50% holdout sample, repeated five times over with a different random seed. The variance of the prediction for the holdout sample is averaged over the five samples. The inclusion of depth, latitude group (West Coast) or longitude group (South Coast), size class and year is considered for the model. Neither latitude group (West Coast) / longitude group (South Coast) or year make any substantial impact on the variance of prediction calculated in this way. Depth and size class do however result in a considerable reduction in the prediction variance. Equation parameters are reported here for calculating the proportion of ***M. paradoxus*** by depth and for each size class.

A model independent comparison between the demersal survey data and the OROP and SADSTIA observer programme data is carried out. This comparison shows that the observer data reflects a very much higher proportion of ***M. paradoxus*** at depths of less than 300 metres than the demersal survey data. Analysis of the data by depth shows that very few trawls are sampled at depths of < 200 metres in the observer data. The demersal survey has a very high proportion of trawls sampled in those depth ranges. Thus a general conclusion that the survey data is a more reliable predictor of species composition in shallow waters seems in order. However, the consistent positive bias in the proportion of ***M. paradoxus*** for the observer data < 300 metres compared to the survey data is a concern. It seems possible that when commercial fishing in the offshore sector occurs at depths < 300 metres, this happens at places where the ***M. paradoxus*** proportion in hake catches is relatively large. This hypothesis needs to be tested in ensuing analyses.

1. Introduction

Gaylard and Bergh (2003, 2004a) used survey data to develop a depth and size based algorithm for splitting hake catches into the component species *M. paradoxus and M.capensis.* In the absence of species information in the commercial catch and effort logbooks, this algorithm has been used since 2004 to produce species-specific catch and CPUE data for input into the stock assessments. Gaylard & Bergh (2007) tested the algorithm using observer data collected by CapFish on behalf of SADSTIA, and found reasonable agreement at an aggregated level.

Gaylard and Bergh (2009) updated the species splitting model based on survey data that had accrued since the work reported by Gaylard and Bergh (2007). They noted that:

- Skipper estimates of size composition appear to be both biased and highly variable estimators of the true size composition when compared to OROP observer data, leading to a resulting under-estimate of the proportion of *M.capensis* of approximately 2 to 3% (of total hake catch) on aggregate, when applied to a typical space/size distribution of commercial catch.

Gaylard and Bergh (2009) consequently proposed an alternative species split model, based on a revised size-classification, as a candidate mechanism for correction of this bias. However, the altered size composition was not adopted by the working group. Consequently this study conforms to use of the following mutually exclusive categories within the specific size ranges:

Large: 58 cm +; Medium: 43 cm to 57 cm; Small: 21 cm to 42 cm

which represents a compromise between the classifications assumed to be in use at the time by I&J and Sea-Harvest, and as reported by Leslie, Rose and Scholte (1998).

2. Methods

2.1 Data acquisition and preparation

This study makes use of survey data for the period 1985 to 2012. The aggregate number of fish sampled per year in the survey data, 1984 – 2012, is shown in Tables 1, and in the SADSTIA observer programme + OROP dataset it is presented in Table 2.

Fig. 1 shows the distribution of depths of demersal survey trawl locations 1984 – 2012, while Fig. 2 shows the distribution of depths of commercial trawls sampled as part of the SADSTIA observer and the OROP programmes. These figures show that the SADSTIA/OROP data which is sampled from commercial trawls shows a very different depth distribution compared to the survey data, where trawls are located on the basis of scientific principles. This has implications for the use of the SADSTIA/OROP data to provide an independent test of the reliability of the species splitting model, since it has less coverage over the mixed range where the species co-occur.

2.2. Definition of the SMALL, MEDIUM and LARGE classification.

The survey and SADSTIA+OROP data have been aggregated by size class to reflect the following size classification

Large: 58 cm +; Medium: 43 cm to 57 cm; Small: 21 cm to 42 cm

2.3 The Model Equation – Survey Data

The survey data are available as the total number of hake sampled for a given cruise and sampling station, with the following ancillary information made available in the data preprocessed for this study: size group: small, medium and large; depth of fishing in metres; calendar year, and for the West Coast, a latitude group, for the East Coast a longitude group.

Both Gaylard and Bergh (2004) and Gaylard and Bergh (2009) used a logistic function to describe the dependence of the proportion of *Merluccius capensis* on depth. The statistical model differed however. In Gaylard and Bergh (2004) the model fit was achieved using an explicit binomial likelihood function. In

addition an 'effective sample size' was fitted in order to address overdispersion in the binomial distribution. ADMB software was used to achieve the model fits.

Gaylard and Bergh (2009) however restructured the data so that existing package software for fitting a logistic regression model could be used to achieve the sigmoid dependence on depth. A shortcoming of that study is that the analysis did not address the overdispersion aspect that Gaylard and Bergh (2004) dealt with via the effective sample size approach.

This study reverts to the methods applied in Gaylard and Bergh (2004). However, whereas Gaylard and Bergh (2004) used a binomial likelihood function explicitly modelled using ADMB software, this document is based on Generalised Linear Models fitted to the data in which the scaling parameter is estimated using the Pearson Chi-squared method. The fits were cross checked between R and SPSS and found to agree exactly.

The model utilises a GLM with a binomial distribution and a logit link function. Model effects are additive in logit space, via an equation of the following form for the West Coast:

$$P = \frac{e^{\Psi}}{1+e^{\Psi}} = \frac{1}{1+e^{-\Psi}} \quad \textbf{(1)}$$

with $$\Psi = \mu + \alpha_y + \phi_{latitude} + \lambda_{sizeclass} + \gamma depth \quad \textbf{(2)}$$

where: P is the proportion of ***Merluccius paradoxus***;

μ	is the intercept;
α_y	is the year parameter for year *y;*
$\phi_{latitude}$	is the latitude parameter*;*
$\lambda_{sizeclass}$	is the size class specific parameter*;*
γ	is the covariate parameter for depth*;*

The same model form is used for the South Coast. However in that case, instead of latitude, the spatial variable is longitude. The above is a basic presentation of the model structure. Modifications to allow for interactions between predictors are not described here, but are considered in the development of the final model.

The advantage of using generalised linear models is that the scale parameter, which is a multiplier on the variance given by the usual binomial mass distribution function, can be fitted. Different values of the scaling parameter do not change the parameter estimates, however they do change the probability levels for different parameter estimates.

2.4 Model Selection

Initial model selection was carried out for the main effects only. The model selection approach was to carry out a 50:50 random split of the available data by station number (representing the individual trawls in the surveys across all surveys). A model was fit on the 'train' dataset and the variance of the residual proportion of *M. paradoxus* was calculated for the 'test' dataset not used in the model fit. Five different random partition were used, the model was refitted to each newly selected 'train' dataset, and the variance of the residual proportions was averaged over these five different random partitions.

3. Results.

The results obtained were as follows:

<u>West Coast</u>

Variance of the proportion of *M. paradoxus* in the full dataset: 0.204

Depth only, variance of residuals of proportions of *M. paradoxus*: 0.117

Depth + Year, variance of residuals of proportions of *M. paradoxus*: 0.116

Depth + Latitude Group, variance of residuals of proportions of *M. paradoxus*: 0.118

Depth + Size Class, variance of residuals of proportions of *M. paradoxus*: 0.047

Depth x Size Class, variance of residuals of proportions of *M. paradoxus*: 0.047

South Coast

Variance of the proportion of *M. paradoxus* in the full dataset: 0.075

Depth only, variance of residuals of proportions of *M. paradoxus*: 0.035

Depth + Year, variance of residuals of proportions of *M. paradoxus*: 0.036

Depth + Longitude Group, variance of residuals of proportions of *M. paradoxus*: 0.034

Depth + Size Class, variance of residuals of proportions of ***M. paradoxus***: 0.020

Depth x Size Class, variance of residuals of proportions of ***M. paradoxus***: 0.018

These results seem unequivocal that the only effects that should be included in the model are depth and size class, for both the West Coast and the South Coast, and that there is no justification for the use of an interaction term between the two.

Note: A different set of random seeds was used for the 50:50 partition of the data into test and train datasets, five times over. The results obtained in this way were very close to the results reported above.

The parameter estimates needed to calculate the predicted proportion of ***M. paradoxus*** are given in Table 3.

South Coast plots: Fig. 3 shows a graphic of the predicted proportions of ***M. paradoxus*** as a function of depth for the three size classes defined, for the South Coast. Fig. 4 shows a plot of the observed proportions corresponding to the predicted values plotted in Fig. 3, again for the South Coast only. Fig. 5 is a plot of the predicted proportions versus the observed proportions. Fig. 6 shows the residual between the observed and predicted number of ***M. paradoxus*** per record in the base data from the surveys, aggregated across the period 1984 to 2012, and plotted against the predicted proportions.

West Coast plots: Fig. 7 shows a graphic of the predicted proportions of ***M. paradoxus*** as a function of depth for the three size classes defined, for the West Coast. Fig. 8 shows a plot of the observed proportions corresponding to the predicted values plotted in Fig. 7, again for the West Coast only. Fig. 9 shows the residual between the observed and predicted number of ***M. paradoxus*** per record in the base data from the surveys, aggregated across the period 1984 to 2012, and plotted against the predicted proportions.

Figs 10 and 11 compare the proportion of ***M. paradoxus*** in 25 metre wide depth bins between the OROP+SADSTIA data and the survey date, for the South Coast.

Figs 12 and 13 compare the proportion of ***M. paradoxus*** in 25 metre wide depth bins between the OROP+SADSTIA data and the survey date, for the West Coast.

Table 4 shows the proportion breakdown of samples by longitude group (1 degree classes) or latitude group (1 degree wide classes) for the South Coast and the West Coast. The summaries in Table 4 are plotted in Fig. 14 (South Coast) and Fig. 15 (West Coast).

4. Discussion

The results seem very clear that neither latitude group (West Coast) / longitude group (South Coast) or year make any substantial impact on the variance of prediction of the species splitting model. Depth and size class do however result in a considerable reduction in the prediction variance.

The model independent comparison between the demersal survey data and the OROP and SADSTIA observer programme data shows that the observer data reflects a very much higher proportion of ***M. paradoxus*** at

depths of less than 300 metres than the demersal survey data. Analysis of the data by depth shows that very few trawls are sampled at depths of < 200 metres in the observer data. The demersal survey has a very high proportion of trawls sampled in those depth ranges. Thus a general conclusion that the survey data is a more reliable predictor of species composition in shallow waters seems in order. However, the consistent positive bias in the proportion of ***M. paradoxus*** for the observer data < 300 metres compared to the survey data is a concern. This disconnect may be due to a number of factors, viz.

1. Misspecification of the length range of hake that correspond to the small, medium and large size catagories used for reporting commercial catches.
2. A different distribution by length in the survey data compared to the commercial data, for the small, medium and large size ranges.
3. Different fishing locations for shallow water trawls in the commercial data compared to the survey data.

These hypotheses need to be tested in ensuing analyses.

5. References.

Gaylard J.D. and M.O. Bergh. 2003. An investigation into the procedure used to split commercial catches of hake on the South African South Coast into *Merluccius paradoxus* and *Merluccius capensis*. BEN/JAN04/SAH2b

Gaylard J.D. and M.O. Bergh .2004a. A size-dependent species splitting mechanism applied to hake catches off the South African West Coast. WG/08/04/D:H:13

Gaylard J.D. and M.O. Bergh M. 2004b. A species splitting mechanism for application to the commercial hake catch data 1978 to 2003. Marine Coastal Management Document WG/09/04/D: H: 21. 8 pp.

Gaylard J.D. and M.O. Bergh. 2007. Further comparison of hake species splits from observer data with the survey-generated splitting algorithm.

Gaylard J.D. and M.O. Bergh. 2009. Update of the hake species split models in the light of more recent survey data and a revision of the large/medium/small size classification. MCM/2009/NOVEMBER/SWG-DEM/…

Leslie R.W. B. Rose and J. Scholte. 1998. Hake grading by Irvin & Johnson and by Sea Harvest. WG/01/98/D:H:03

6. Acknowledgements.

We thank Tracey Fairweather for the provision of the relevant data upon which this study is based.

Tables

Table 1. Total number of hake sampled during demersal surveys in South African waters, 1984 – 2012.

Year	Number of Fish	Year	Number of Fish
1984	36634	1999	49940
1985	45256	2000	42716
1986	55392	2001	71564
1987	57307	2002	86078
1988	58270	2003	97686
1989	55724	2004	146278
1990	87461	2005	41136
1991	68535	2006	89471
1992	60586	2007	78591
1993	70537	2008	63385
1994	74648	2009	61407
1995	87845	2010	47617
1996	53975	2011	45899
1997	52077	2012	33899

Table 2. Total number of hake sampled during commercial trawling operations in South African waters, 2002 – 2012, as part of the OROP and SADSTIA Observer programmes combined.

Year	Number of Fish
2002	121858
2003	653475
2004	520439
2005	469139
2006	864232
2007	465220
2008	617855
2009	622294
2010	381736
2011	200010
2012	110636

Table 3. Model parameter estimates for the West and South Coasts for the proportion of *M. paradoxus*, using the preferred version of the model where the main effects are depth and size class.

	West Coast	South Coast
μ	-12.851	-23.183
λ_{small}	5.788	10.997
λ_{medium}	2.049	7.391
λ_{large}	0.000	0.000
γ	0.037	0.073

Table 4. Proportion breakdown of samples by longitude group (1 degree classes) or latitude group (1 degree classes) for the South Coast and West Coast respectively. See plots in Figs 14 and 15.

	Survey		OROP+SADSTIA	
	South Coast		South Coast	
Longitude	All	<300 metres	All	<300 metres
21	48.35	49.47	14.9	23.21
22	15.92	15.88	3.69	3.97
23	13.56	13.67	6.28	9.14
24	7.03	6.63	28.71	25.1
25	6.87	5.83	44.36	31.73
26	8.21	8.47	2.06	6.86
27	0.06	0.06	0	0

	Survey		OROP+SADSTIA	
	West Coast		West Coast	
Latitude	All	<300 metres	All	<300 metres
29	14.48	17.28	0.01	0.03
30	15.26	17.68	2.89	4.82
31	16.42	14.03	5.06	3.49
32	17.65	13.13	10.22	6.01
33	11.97	13.23	21.92	7.55
34	13.5	11.12	28.1	14.59
35	10.24	12.93	16.52	24.5
36	0.48	0.6	15.28	39.01

Figures

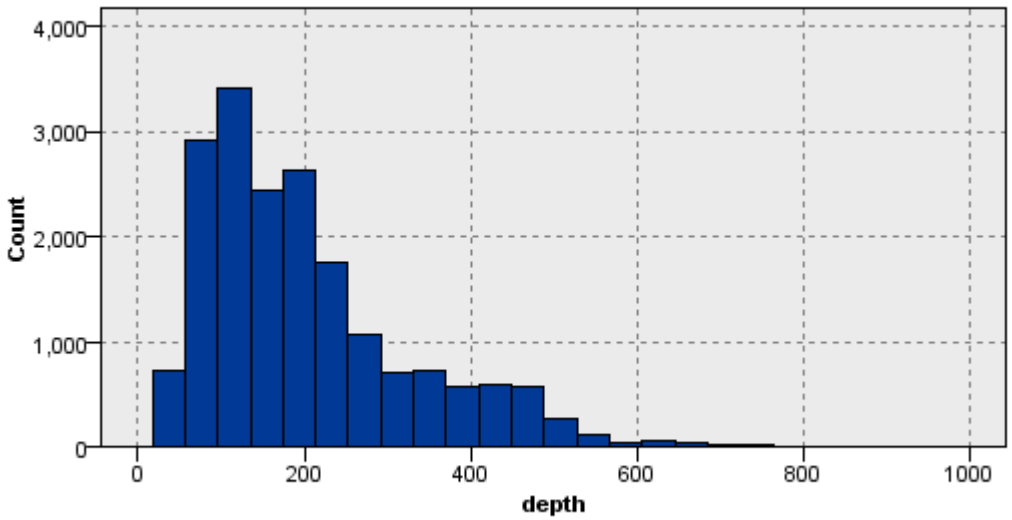


Figure 1. The distribution of depths of demersal survey trawl locations 1984 – 2012.

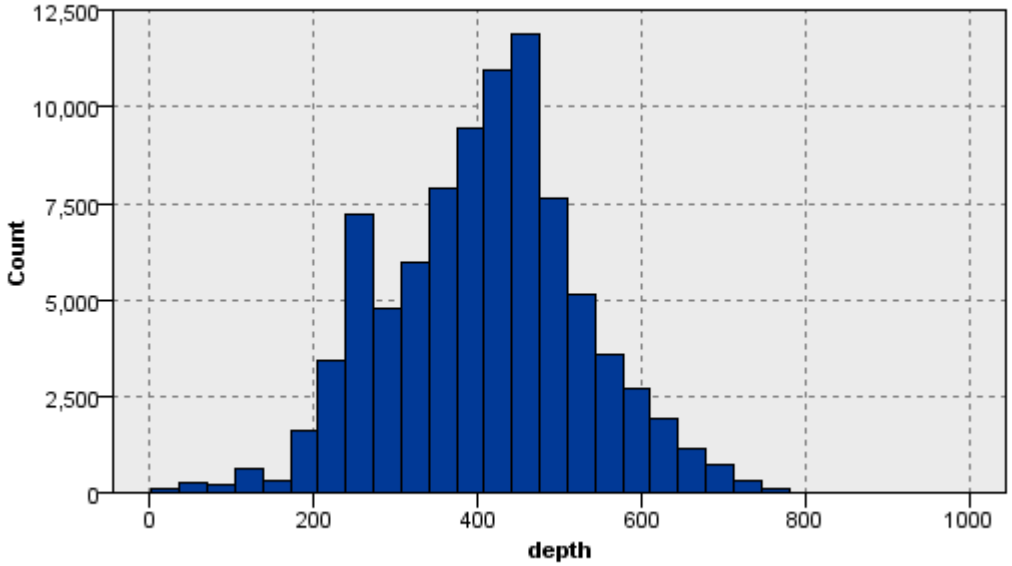
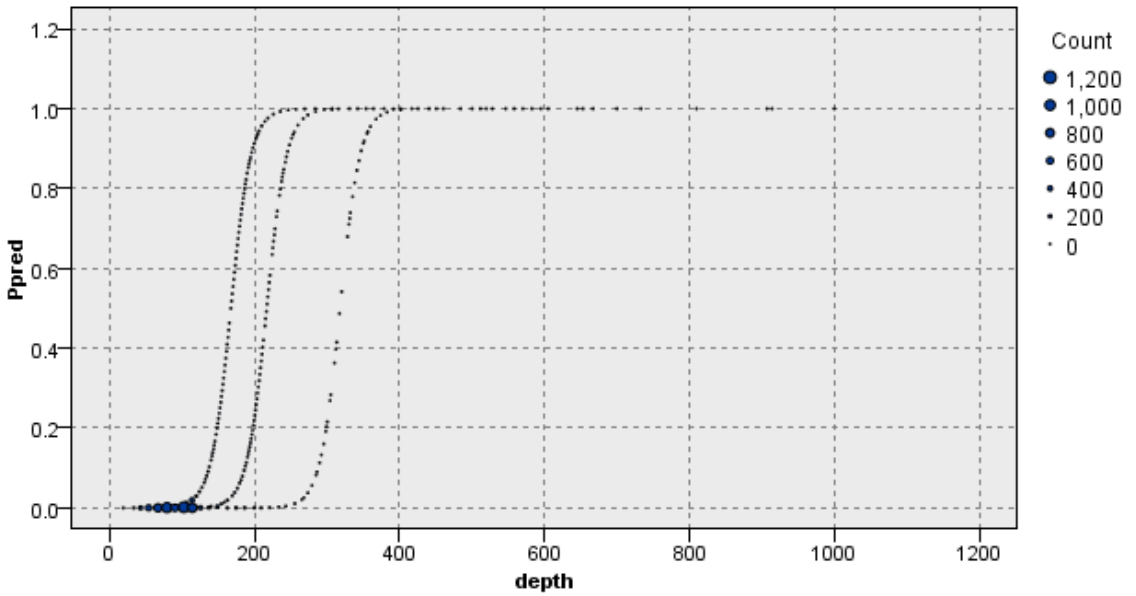


Figure 2. The distribution of depths of commercial trawls sampled as part of the SADSTIA observer and the OROP programmes.



173

Figure 3. A plot of predicted proportions of *M. paradoxus*, versus depth, for all records for the South Coast in the survey database for the period 1984 – 2012.

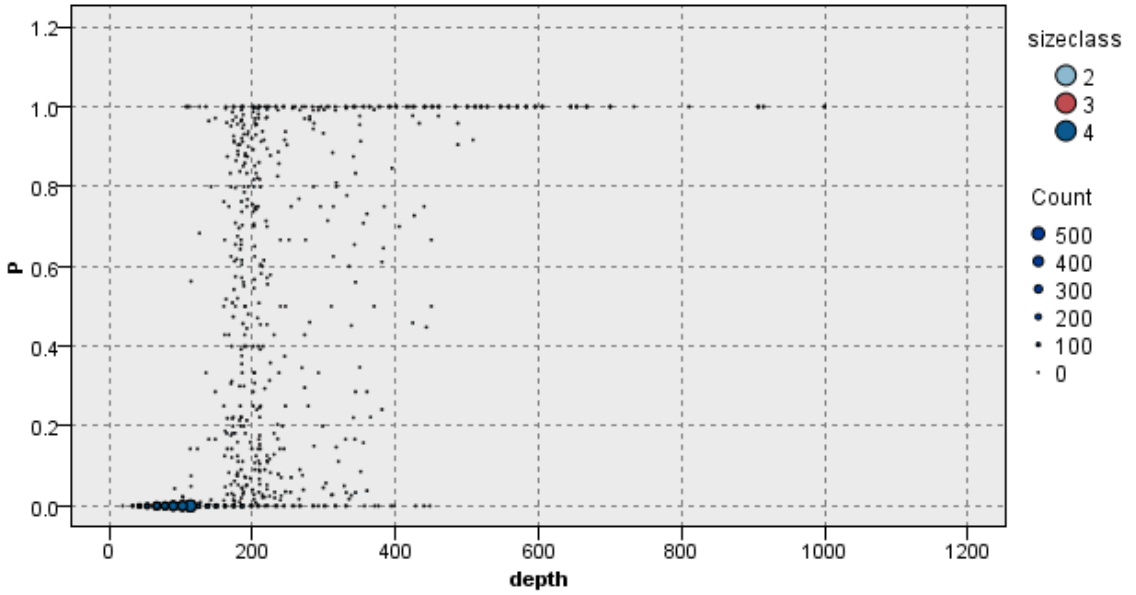


Figure 4. A plot of observed proportions of *M. paradoxus*, versus depth, for all records for the South Coast in the survey database for the period 1984 – 2012.

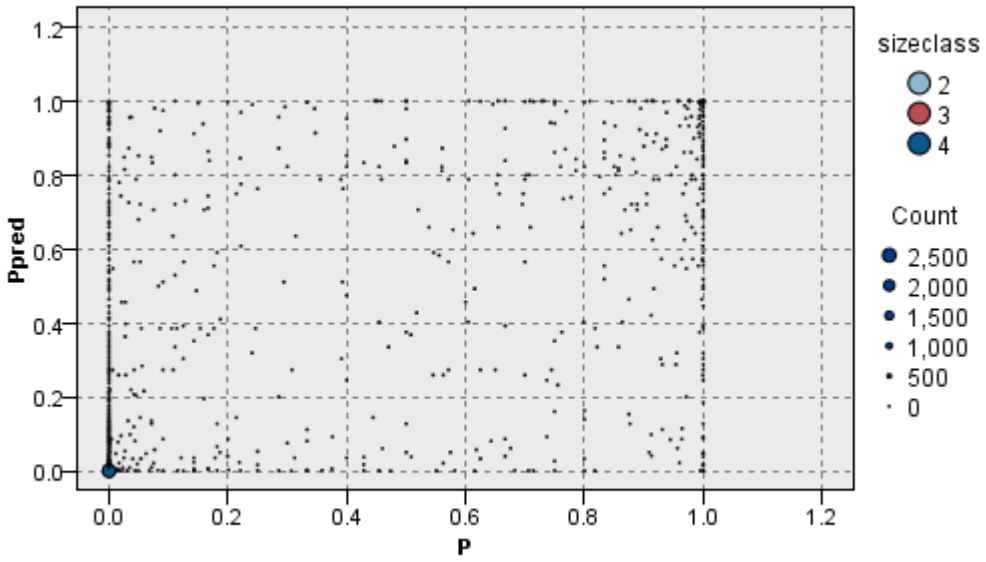


Figure 5. A plot of observed proportions of *M. paradoxus*, versus model predicted proportions, for the preferred version of the model in which the main effects are size class and depth, for all records for the South Coast in the survey database for the period 1984 – 2012.

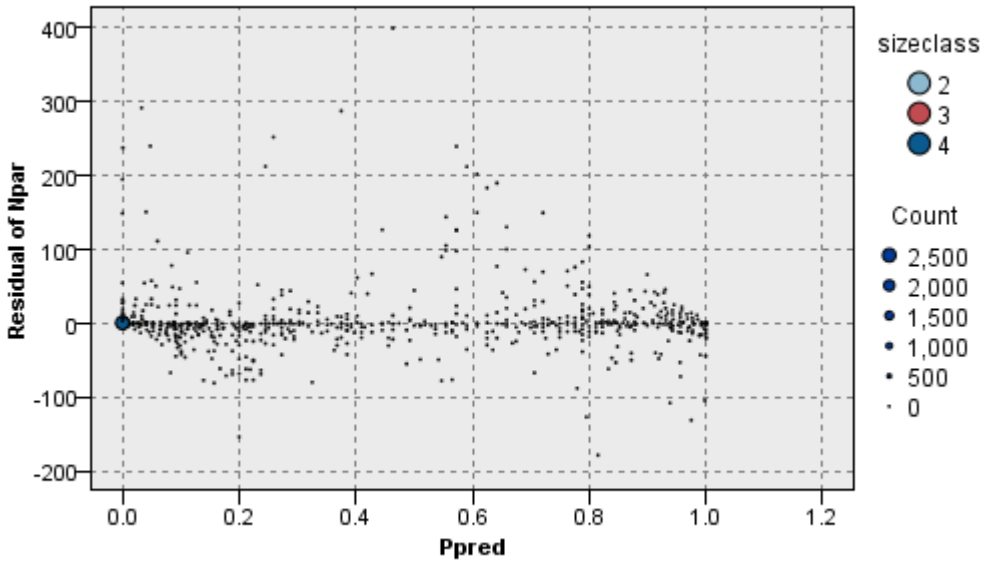


Figure 6. A plot of the residual between observed and predicted number of M. paradoxus per record in the trawl dataset for the South Coast, (for the model using depth and size class as sole main effects) 1984 – 2012.

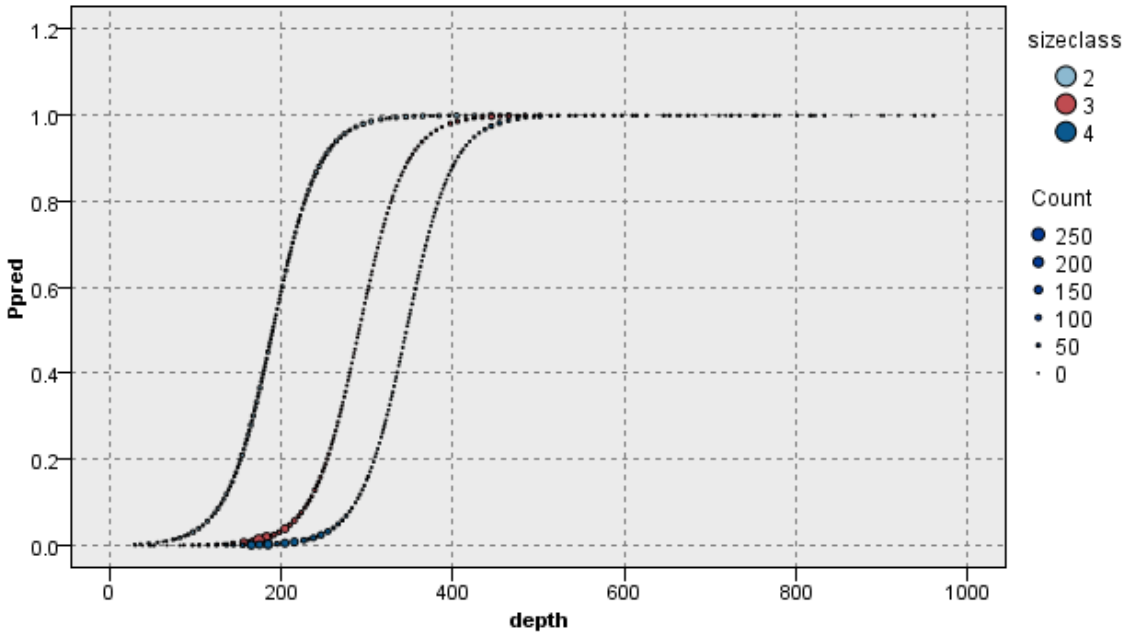


Figure 7. A plot of predicted proportions of *M. paradoxus*, versus depth, for all records for the West Coast in the survey database for the period 1984 – 2012.

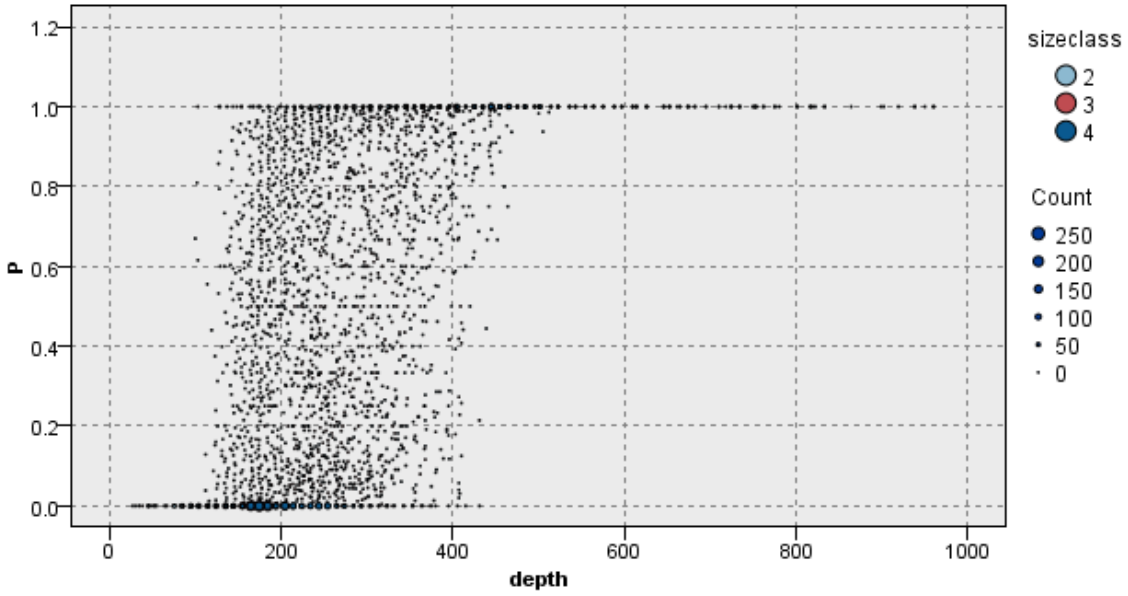


Figure 8. A plot of observed proportions of *M. paradoxus*, versus depth, for all records for the West Coast in the survey database for the period 1984 – 2012.

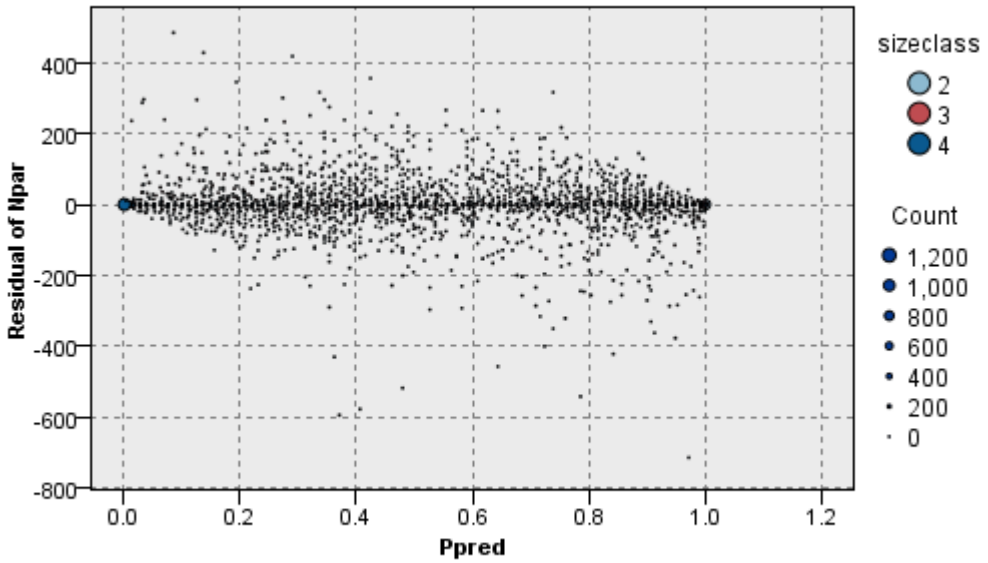


Figure 9. A plot of the residual between observed and predicted number of *M. paradoxus* per record in the trawl dataset for the West Coast, (for the model using depth and size class as sole main effects) 1984 – 2012.

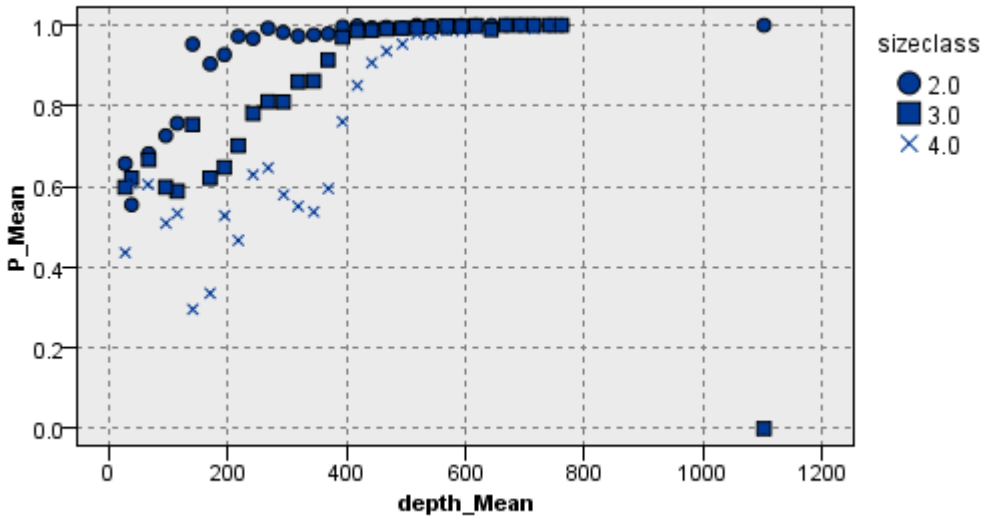


Figure 10. The average proportion of M. paradoxus for different depths (aggregated by depth bins of width 25 metres) in the OROP+SADSTIA dataset for the South Coast.

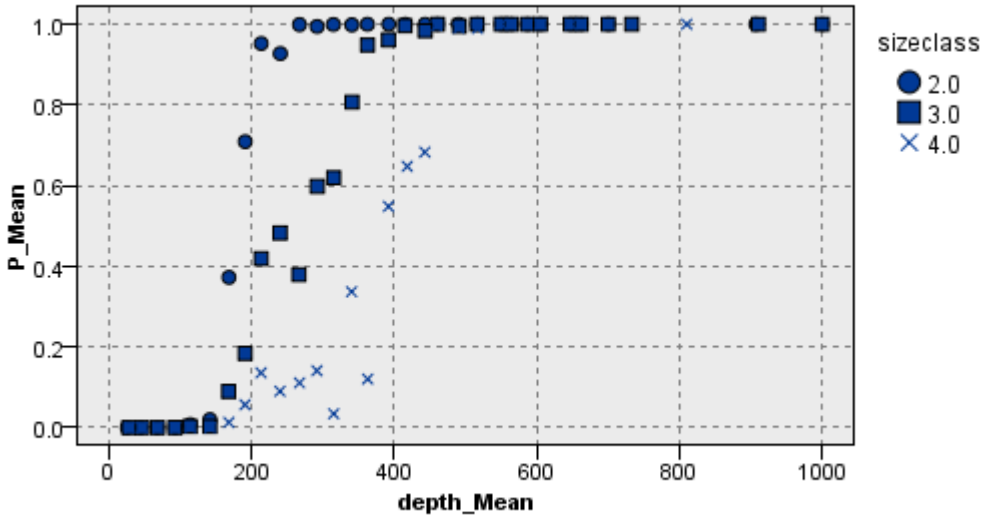


Figure 11. The average proportion of *M. paradoxus* at different depths (aggregated by depth bins of width 25 metres) in the demersal survey data for the South Coast.

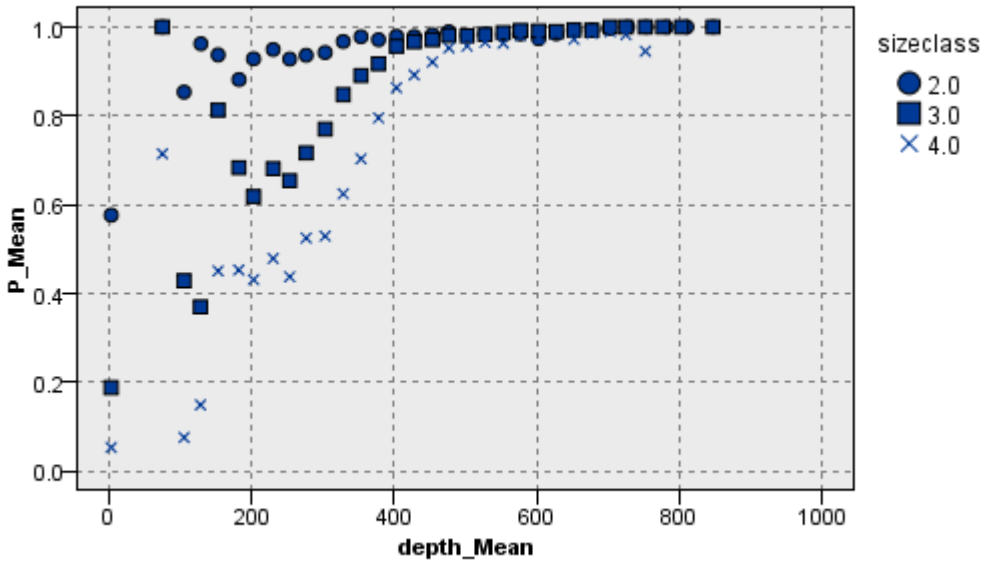


Figure 12. The average proportion of M. paradoxus for different depths (aggregated by depth bins of width 25 metres) in the OROP+SADSTIA dataset for the West Coast.

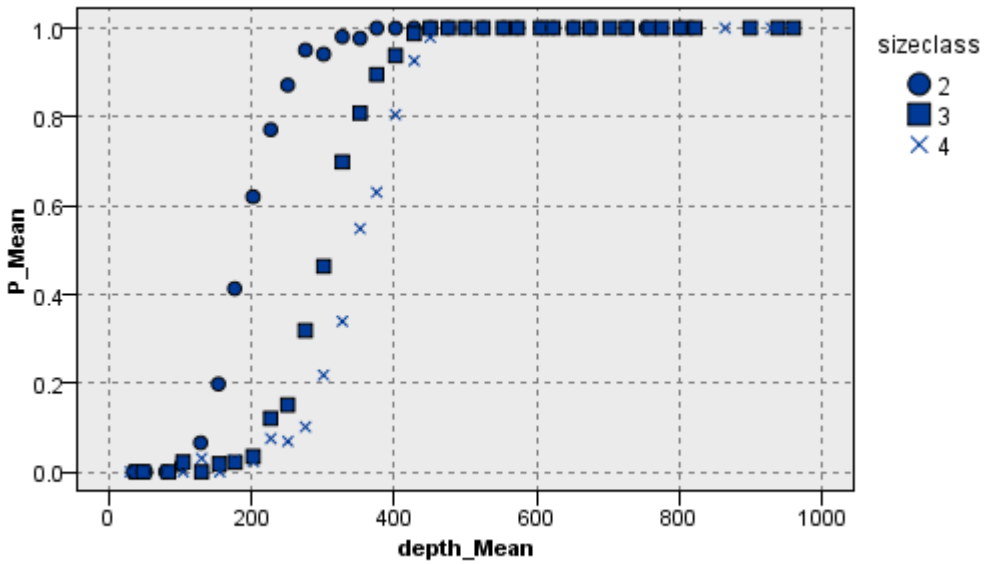


Figure 13. The average proportion of M. paradoxus for different depths (aggregated by depth bins of width 25 metres) in the demersal trawl dataset for the West Coast.

Figure 14. South Coast proportion of samples by longitude degree, all samples left panel, < 300 metres samples right panel.

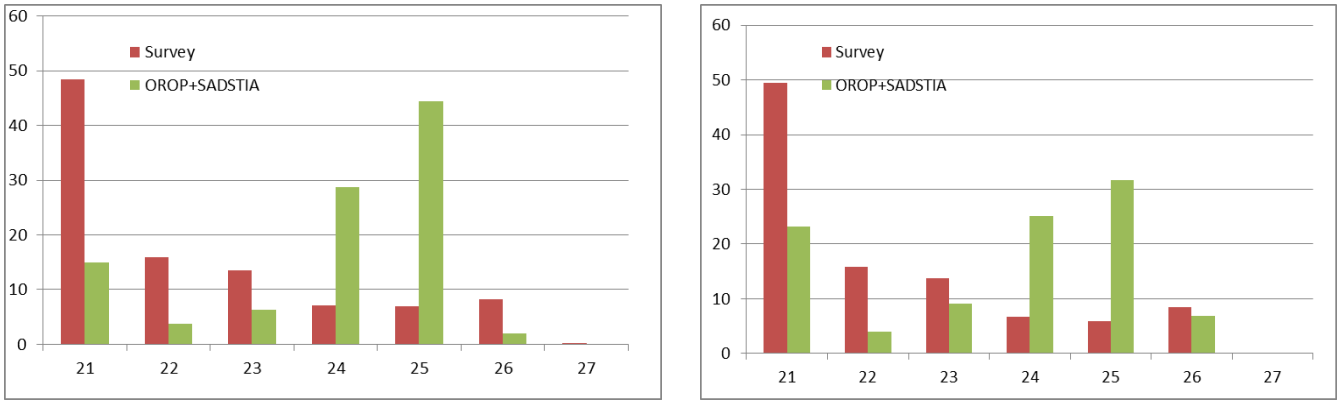


Figure 15. West Coast proportion of samples by latitude degree, all samples left panel, < 300 metres samples right panel.