Initial results from using a GLMM to standardise the commercial CPUE data for Namibian hake, and an update of the GLM analyses to incorporate the extra data now available for 2004 and 2005

Anabela Brandão and Doug S. Butterworth

Marine Resource Assessment and Management Group (MARAM)
Department of Mathematics and Applied Mathematics
University of Cape Town
Rondebosch, 7701, South Africa

March 2006

Abstract

Previous analyses to standardise the commercial CPUE hake data have encountered problems related to the presence of a significant year-vessel interaction effect in the GLM analyses. These include the fact that the CPUE data do not represent a balanced design and also that the number of parameters to be estimated exceeds the capabilities of statistical packages available. Furthermore, the presence of latitude or depth interactions with year in the GLM analyses complicates the standardisation of the CPUE series as integration over area becomes necessary, and procedures are needed to deal with missing cells. GLMM methodology is applied to overcome these problems. The conventional GLM analysis to standardise the CPUE for hake is applied to updated data for the period 1992 to 2005. The impact of a grid sorter on vessels is taken into account. The GLM standardised CPUE index shows a downward trend from 1999. An upward trend is evident from 2002 until 2004, but there is a 17% decrease from 2004 to 2005. Grid sorters are estimated to decrease catching efficiency by some 3%. GLMM standardised CPUE series show similar trends to those for the GLM, though the GLMM standardised series reflect a slightly larger decrease over the full period considered. The GLMM that includes all year interactions (i.e. including that with vessel) as random effects is selected as the best model of those considered in terms of both the AIC criterion and deviance analyses.

Introduction

Previous GLM analyses of the commercial CPUE data for Namibian hake have shown that there is a significant interaction between vessels and year (Brandão *et al*. 2001, Brandão and Butterworth 2001). However, inclusion of this interaction in the GLM (General Linear Model) analyses used to standardise the hake CPUE data is problematic in two respects. First, the number of parameters to be estimated in the GLM is too many for available statistical packages to handle. Secondly, the hake CPUE data are not representative of a balanced design as new vessels enter and old vessels leave the fishery, causing the number of records available by year and vessel to be very "spotty" with empty cells, so that the approach applied to date has had to use GRT as a surrogate factor for vessel. This paper investigates the use of a GLMM (General Linear Mixed Model) approach to take account of year-vessel interactions by treating these as a random effect. Note that this then explicitly accounts for differences between wetfish and freezer trawlers.

The basecase GLM used to standardise the commercial hake CPUE data includes several terms with interactions with year (which imply changing spatio-temporal distribution patterns) (Brandão and Butterworth 2005). To obtain a standardised CPUE series to be used as an index of relative abundance when input to assessment models, the CPUE itself is assumed to be proportional to local density, so that integration over area (and, conventionally, averaging over month) is necessary to provide a quantity proportional to overall abundance. This integration/averaging is unnecessary in the absence of such interactions, because then the exp(β_{year}) term alone will then be proportional to abundance. A GLMM is therefore also investigated in which all interactions with year are treated as random effects, overcoming the need to integrate/average over area and month to obtain a standardised CPUE series for hake.

This paper also applies the "conventional" GLM model as implemented in Brandão and Butterworth (2005), which incorporates information given by some operators that they had used a grid sorter on their vessels in 2003, to provide an updated index of abundance for the Namibian hake resource that includes the extra data now available for 2004 and 2005. However, as these extra data became available only very recently, there was time to perform the GLMM analyses only on the data used in Brandão and Butterworth (2005), for which the 2004 data are incomplete.

Basecase GLM to standardise the CPUE

The GLM of Brandão and Butterworth (2005), which allows for possible annual differences in hake areal distribution and accounts for the impact of a grid sorter, is used to standardise the commercial hake CPUE data. This model is given by:

$$\ln(CPUE) = \mu + \alpha_{lat} + \beta_{year} + \gamma_{month} + \lambda_{depth} + \omega_{GRT} + \varphi grid + \eta_{year\times lat} + \theta_{year\times depth} + \tau_{depth\times lat} + \kappa_{month\times lat} + \zeta_{year\times month} \quad (1)$$

where:

μ	is the intercept,
lat	is a factor with 13 levels of degrees latitude (17°S–29°S),
year	is a factor with 14 levels associated with the years 1992–2005,
month	is a factor with 12 levels (January– December),
depth	is a factor with 4 levels ("200" for depths ≤ 299 m, "300" for 300–399 m, "400" for 400–499 m and "500" for ≥ 500 m),
GRT	is a factor with 25 levels associated with increments of 100 of the gross tonnage of a vessel; this commences at "0" for tonnages in the range 0–99, continuing to "1600" for 1600–1699, "1700" for 1700–1899 (as there were too few observations with tonnages in the 1800s), and "1900" for 1900–1999, whereafter there are five further categories 2200–2299, 3000–3099, 3100–3199, 3800–3899, 3900–3999, and 5600–5699.
grid	represents a Boolean variable with a value of "0" when no grid sorter was reported and "1", when a grid sorter was reported to have been implemented.
year×*lat*	is the interaction between year and latitude,
year×*depth*	is the interaction between year and depth,
depth×*lat*	is the interaction between depth and latitude,
month×*lat*	is the interaction between month and latitude,
year×*month*	is the interaction between year and month, and
ε	is the error term assumed to be normally distributed.

For this model, because of interactions with year (which imply changing spatio-temporal distribution patterns), the standardised CPUE series is obtained from:

$$CPUE_{year} = \left[\sum_{strata} \sum_{month} \left(\exp\left(\begin{matrix} \mu + \alpha_{lat} + \beta_{year} + \gamma_{month} + \lambda_{depth} + \overline{\omega} + \eta_{year\times lat} + \\ \theta_{year\times depth} + \tau_{lat\times depth} + \kappa_{month\times lat} + \zeta_{year\times month} \end{matrix} \right) * A_{stratum} \right) \right] / 12 \quad (2)$$

where:

$\overline{\omega} = \omega_{GRT=700}$, and

$A_{stratum}$ is the surface area of the stratum defined by the degree of latitude and depth range concerned.

Often models with interaction terms have missing cells for certain combinations of levels of factors. To be able to compute equation (2) for standardising the CPUE, the missing cells were replaced by the average of the estimable factors that "surround" the missing cell (that is, factors from the cells "above" and "below" and the factors from the cells on either "side" (thinking in terms of a map)). When there are missing cells adjacent to each other, the procedure is more complicated in that an order in which missing cells are filled must be specified as the results are not order-invariant. An example of the procedure used to replace missing cells with the average of other cells is shown in more detail in Brandão *et al.* (2001).

The sizes of the areas for each stratum, based on values used for the *Nansen* surveys, are given in Table 1. These strata do not correspond exactly to those chosen for the GLM. Level 200 for depth in the GLM includes all depths less than or equal to 299 m, because there were too few data points in the < 200 m range to make this a separate level for the depth factor. Similarly, the depth level of 500 m in the GLM includes all data points with depths greater than or equal to 500, and this includes a number of trawls at depths > 600 m. However, since the vast majority of fishing takes place between depths of 200–600 m, use of areas as listed in Table 1 would seem to reflect a reasonable approximation to the fished component of the resource.

GLMM to standardise the CPUE

The GLMM approach applied treats the interactions with year as random effects. Thus the model implemented has the form:

$$\ln(CPUE) = \mathbf{X}\alpha + \mathbf{Z}\beta + \varepsilon \qquad (3)$$

where

α is the unknown vector of fixed effects parameters,

$\mathbf{X}$ is the design matrix for the fixed effects,

β is the unknown vector of random effects parameters,

$\mathbf{Z}$ is the design matrix for the random effects, and

ε is an error term assumed to be normally distributed and independent of the random effects.

This approach assumes that both the random effects and the error term have zero mean, i.e. $E(\beta) = E(\varepsilon) = 0$, so that $E(\ln(CPUE)) = \mathbf{X}\alpha$. The variance-covariance matrix for the residual errors

(ε) is denoted by **R** and that for the random effects (β) by **G**. The analyses undertaken here assume that the residual errors as well as the random effects are homoscedastic and uncorrelated, so that both **R** and **G** are diagonal matrices given by:

$$\mathbf{R} = \sigma_{\varepsilon}^2 \mathbf{I}$$
$$\mathbf{G} = \sigma_{\beta}^2 \mathbf{I}$$

where **I** denotes an identity matrix. Thus, in the mixed model, the variance-covariance matrix (**V**) for the response variable is given by:

$$\text{Cov}(\mathit{Incr}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{R},$$

where $\mathbf{Z}^{\mathrm{T}}$ denotes the transpose of the matrix **Z**.

The estimation of the variance components (**R** and **G**), the fixed effects (α) and the random effects (β) parameters in GLMM requires two steps. First the variance components are estimated by the method of residual maximum likelihood (REML) (Patterson and Thompson, 1971), which produces unbiased estimates for the variance components as it takes into account the degrees of freedom used in estimating the fixed effects. The REML method maximises the likelihood of a set of error contrasts rather than the likelihood of the whole data set. These error contrasts are linear combinations of the observations and have an expectation of zero[1]. The residual log-likelihood is given by:

$$-2\ln L(\mathbf{R},\mathbf{G};\ln(\mathit{CPUE})) = (n-p)\ln(2\pi) - \ln\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right| + \ln\left|\mathbf{V}\right| + \ln\left|\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right| + \ln(\mathit{CPUE})^{\mathrm{T}}\mathbf{P}\ln(\mathit{CPUE}) \qquad (5)$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$, *n* is the number of data points, *p* is the degrees of freedom used in estimating the fixed effects and the minus sign in the superscript denotes a generalised inverse of the matrix concerned. Once estimates of **R** and **G** ($\hat{\mathbf{R}}$ and $\hat{\mathbf{G}}$) have been obtained, generalised least squares estimates for the fixed effects parameters (α) can be obtained from:

$$\hat{\alpha} = \left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\ln(\mathit{CPUE}) \qquad (6)$$

and predictors for the random effects parameters (β) are obtained from the best linear unbiased predictors (BLUPs) given by:

$$\hat{\beta} = \hat{\mathbf{G}}\mathbf{Z}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\left(\ln(\mathit{CPUE}) - \mathbf{X}\hat{\alpha}\right) \qquad (7)$$

[1] An error contrast is defined to be a linear combination $\mathbf{K}\ln(\mathit{CPUE})$ of the observations such that $\mathrm{E}\left(\mathbf{K}\ln(\mathit{CPUE})\right) = 0$, i.e. $\boldsymbol{KX} = 0$. One such choice is given by $\mathbf{K} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$, so that $\mathbf{K}\ln(\mathit{CPUE}) = \left(\ln(\mathit{CPUE}) - \mathbf{X}\tilde{\alpha}\right)$, where $\tilde{\alpha}$ is the ordinary least squares estimate of α, i.e. the REML method maximises the likelihood of the ordinary least squares residuals $\left(\ln(\mathit{CPUE}) - \mathbf{X}\tilde{\alpha}\right)$, rather than the likelihood of $\ln(\mathit{CPUE})$.

GLMMs investigated

Three GLMMs were applied to the hake CPUE data. Initially, only a year-vessel interaction was considered as a random effect. In this case, the vector of fixed effects parameters (α) includes all the parameters of equation (1) above, except that the factor GRT has been replaced with the individual vessels (228 in total), and the random effects parameters (β) consist of the year-vessel interactions. A second GLMM treated all year interactions included in the GLM as random effects (i.e. terms in equation (1) were taken to be fixed effects, unless they involved an interaction with year in which case they were taken to be a random effects. In this model the GRT factor was again replaced by a vessel factor. The last GLMM considered added a year-vessel interaction as a further random effect to those included in the GLMM just described. Computations were effected using the statistical package GenStat 8.1.

Model Implementation

Commercial tow information for the years 1992 to December 2005 has been used for the GLM analyses. A total of 161 725 data points (vessel-days of fishing) was available for the analyses. For the GLMM analyses, data for the years 1992 to November 2004 as available for the previous year's analyses has been used. Note must be taken that, although for the previous analyses data were available until November 2004, the extra data made available this year for 2004 included more than only those for an extra month (December 2004) of fishing. For the same period considered last year (i.e. January to November 2004) the number of tows recorded increased from 41 583 to 55 643. Also, because of a problem with the extraction of the data on a daily basis, tow-by-tow data were used to calculate the data aggregated over a day used in the present analyses (as past GLM analyses have used data extracted on that basis). However, the catches recorded on a tow-by-tow basis reflect the captains' estimates, as the landed catch data are available only on a daily basis. Therefore, for 2004 and 2005, the catches used are captains' estimates, while for the earlier years they are the landed catches (as per past practice).

In the GLMM analyses the GRT factor was replaced by a vessel factor. However, some vessels have two codes in the database and as information linking multiple codes used for the same vessel was provided only at a very late stage of these analyses, this linkage has not been taken into account in the data used (i.e. a single vessel is in some cases treated as two different vessels). Furthermore, as the most recent data for 2004 and 2005 were not made available in time for these GLMM analyses, the results shown should be seen as preliminary.

The same approach of Voges (2000) was followed to categorise variables to use as factors in the GLM analyses. The one exception is that the depth variable has been considered in this paper to be a categorical variable with levels representing (effectively) every 100 m depth range.

In 2003, not all operators furnished information on whether grid sorters were used on their vessels. In the absence of this information, unless an operator specifically recorded that a grid sorter was used, it was assumed that no grid sorter was used. From January 2004 all vessels are assumed to use a grid sorter.

Results and Discussion

The basecase GLM model accounts for 43.5% of the total variation of hake CPUE. Table 2 provides standardised CPUE values derived from the basecase GLM. For comparison, the standardised CPUE values obtained previously with commercial CPUE data as earlier provided up to November 2004 (Brandão and Butterworth 2005) are also shown. Figure 1 shows the index of abundance provided by this approach. This is compared to the index of abundance from the previous year's GLM analysis. Both indices have been normalised to their mean over the first thirteen years. The standardised CPUE abundance indices show a downward trend since 1999. However since 2002 there has been an upward trend in the CPUE indices, with a 17% increase from 2003 to 2004, though the 2005 value is 17% down from that for 2004. The grid sorter is estimated to decrease catch rates by about 3%.

Table 3 provides standardised CPUE values as estimated using the three different GLMMs considered. For comparison, the standardised CPUE values obtained from a GLM analysis for the same data are also shown (Brandão and Butterworth 2005). Figure 2 shows the indices of abundance provided by the random effects models. These are compared to the index of abundance from the basecase GLM of the previous year's analyses. All indices have been normalised to their mean over the thirteen years considered. All the indices show similar trends, though those obtained from a GLMM standardisation reflect a slightly greater decrease over the whole period (Table 3 and Fig. 2).

Table 4 shows the deviance (given as -2*log-likelihood) and the Akaike's Information Criteria (AIC) for each of the GLMMs fitted. The reduction in the deviance between two nested models is identical to the log-likelihood ratio statistic for testing the hypothesis of the comparison of the two models which is approximately χ^2 distributed with degrees of freedom corresponding to the difference in the number of parameters of the two models. The GLMM which treats all year

interactions (including a year-vessel interaction) as random effects has both the lowest AIC value and shows a statistically significant improvement compared to the other models.

Future work

The additional data for 2004 and 2005 used in this analysis became available only very recently. Hence there are several issues that need to be pursued and/or further examined in due course:

- Obtain the data on catches for 2004 and 2005 made daily by each vessel that are compatible to those for other years (i.e. landed catches)
- Increase the number of depth levels for deeper depths, as in more recent years more tows have taken place at these greater depths than at the beginning of the present commercial CPUE series.
- Obtain the ocean area for each degree latitude for depths > 600 m.
- Obtain clarification on which vessels have two codes in the database for previous years, and how these link.
- Examine the tow-by-tow CPUE data to determine whether the present criteria for aggregating CPUE to provide daily data (i.e. the recorded depth and position of the aggregated daily tow is the information for the first tow that took place on that day) is appropriate.
- Check that the random effects estimated in the GLMM models do not show systematic patterns (i.e. are consistent with the assumption of randomness).

Acknowledgements

Data for this study have been provided by NatMIRC. Funding from the Namibian Hake Association is gratefully acknowledged.

References

Brandão, A. and Butterworth, D.S. 2001. Effect of incorporation of vessel factors in the GLM analyses of the CPUE data for Namibian hake. BENEFIT Workshop Document: BEN/NOV01/NH/1b: 1–9.

Brandão, A. and Butterworth, D.S. 2005. Updated standardisation of commercial CPUE data of Namibian hake for the period 1992 to 2004. Namibian Ministry of Fisheries and Marine Resources document: HWG/Wkshop/2004/03/Doc2.

Brandão, A., Butterworth, D.S., and Voges, L. 2001. An updated application of GLM analyses to the CPUE data for Namibian hake. Namibian Ministry of Fisheries and Marine Resources document: HWG/Wkshop/2001/Doc4.

Voges, L. 2000. Investigation of Namibian hake catch statistics using a multiplicative model. BENEFIT Workshop Document: BEN/NOV00/NH/2b.

Table 1. The ocean area (nm^2) of each stratum defined by each latitude and depth range. Note that a latitude indicated as, for example, 18° refers to the latitudinal range from 18° to 19°.

	Depth (m)			
Latitude (S)	**201-300**	**301-400**	**401-500**	**501-600**
17°	228	66	60	71
18°	796	144	133	144
19°	972	1042	301	306
20°	889	947	262	297
21°	609	863	216	252
22°	1142	765	152	128
23°	1074	647	230	179
24°	866	695	218	156
25°	1053	553	219	153
26°	953	1444	687	145
27°	498	853	483	168
28°	410	154	141	158
29°	452	345	285	126

Table 2. Standardised CPUE series (each normalised to their mean over the first thirteen years considered) obtained by fitting a General Linear Model (GLM) to the observed CPUE data for Namibian hake. For both standardised CPUE series, the impact of a grid sorter has been taken into account.

Year	Previous year	This year
1992	1.713	1.711
1993	1.859	1.856
1994	1.404	1.393
1995	0.910	0.915
1996	0.789	0.782
1997	0.866	0.863
1998	1.114	1.126
1999	1.129	1.134
2000	0.782	0.779
2001	0.611	0.617
2002	0.490	0.495
2003	0.615	0.615
2004	0.717	0.717
2005	—	0.598

Table 3. Standardised CPUE series (each normalised to their mean over the thirteen years considered) obtained by fitting several General Linear Mixed Models (GLMM) to the observed CPUE data for Namibian hake. For comparison, the GLM standardised CPUE series obtained for the same data is also given. For all these standardised CPUE series, the impact of a grid sorter has been taken into account.

Year	GLM	GLMM		
		Year-vessel interaction only	All year interactions (but not with vessel)	All year interactions (including vessel)
1992	1.713	1.849	1.876	1.902
1993	1.859	2.030	2.043	2.248
1994	1.404	1.412	1.408	1.346
1995	0.910	0.871	0.963	0.892
1996	0.789	0.752	0.740	0.700
1997	0.866	0.841	0.846	0.853
1998	1.114	1.093	1.042	1.060
1999	1.129	1.116	1.052	1.041
2000	0.782	0.756	0.758	0.738
2001	0.611	0.571	0.595	0.582
2002	0.490	0.475	0.473	0.465
2003	0.615	0.575	0.542	0.546
2004	0.717	0.661	0.662	0.627

Table 4. Deviance (given as -2*log-likelihood), number of parameters and the Akaike's Information Criterion (AIC) value for each of the GLMMs fitted. The reduction in the deviance between two nested models is identical to the log-likelihood ratio statistic for testing the hypothesis of a statistically significant improvement between two models. The difference in the AIC between the basecase GLM and the different GLMMs fitted (∆AIC) is also shown. The model with the lowest AIC and the model with the lowest statistically significant deviance is shown italicised in bold.

Model	Deviance	Number of parameters	AIC	∆AIC
GLM	45 920	542	47 004	
GLMM: Year-vessel interaction only	25 751	748	27 247	19 757
GLMM: All year interactions (but not with vessel)	36 625	440	37 505	9 499
GLMM: All year interactions (including vessel)	***25 107***	441	***25 989***	21 015

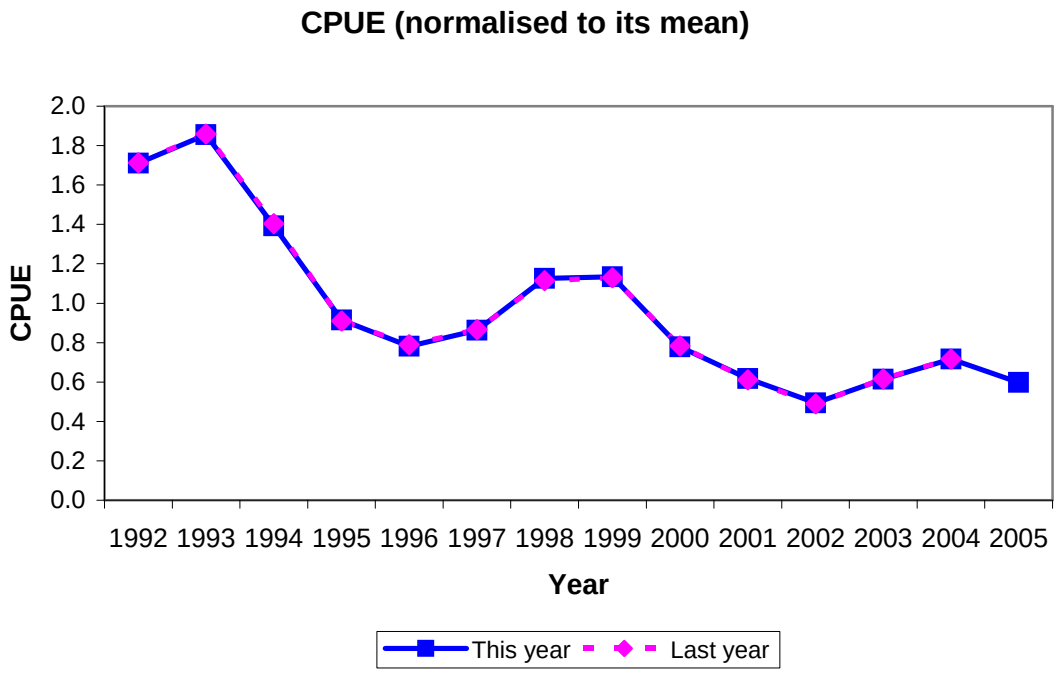

Figure 1. Index of abundance (normalised to its mean over the first thirteen year period) for Namibian hake obtained from fitting the GLM model. For comparison the standardised CPUE series (also normalised to its mean over the thirteen year period) obtained when the model was fitted to the CPUE data as previously available to November 2004 is also shown, though differences are indistinguishable at this scale of plot.

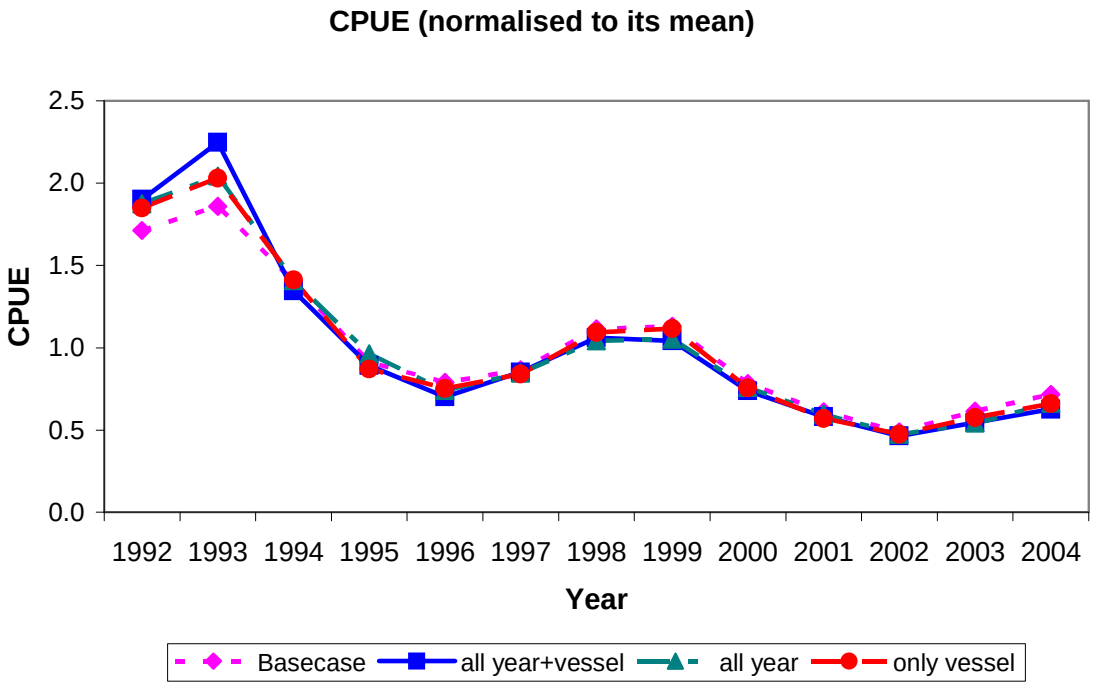

Figure 2. Index of abundance (normalised to its mean over the thirteen year period) for Namibian hake obtained from fitting the different GLMMs. For comparison the standardised CPUE series (also normalised to its mean over the thirteen year period) obtained when the basecase GLM was fitted to the CPUE data is also shown.