GLM and GLMM standardisation of the commercial abalone CPUE for Zones A-D

Anabela Brandão and Doug S. Butterworth

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

August 2012

Abstract

This paper presents an update of the standardisation of the abalone CPUE using the conventional GLM approach and adding new data for the 2010/2011 fishing in Zones A and B. A GLMM approach to standardising abalone CPUE data is also investigated.

Introduction

The catch-per-unit-effort (CPUE) General Linear Model (GLM) standardisation procedure described in Plagányi and Edwards (2007) and Brandão and Butterworth (2009) has been applied to the commercial abalone data for Zones A-D to incorporate further data now available for Zones A and B for the 2011 Model-year, where a Model-year *y* runs from October of year *y*-1 to September of year *y*. The principle objective of the GLM analysis is to obtain series of relative abundance indices that have been standardised by incorporating important covariates in the explanation of abalone CPUE variation.

In this paper a General Linear Mixed Model (GLMM) approach to obtain standardised commercial CPUE series for Zones A-D is also investigated.

The data

Commercial catch data (as kg whole mass), and effort data (as total duration of dives in minutes for each day dived) are available for the period 1980 to 2011. The covariates included in the GLM analysis include the date (in terms of Model-year and season (3-monthly periods)), the divers, and the Zones that were dived. Zone C is split into subareas CNP (nonpoached) and CP (poached). Records with a dive time less than 10 minutes were excluded as well as outliers based upon observations with large residuals (> 6 standard deviations) in an initial GLM fit. Years which had too few records (less than eight) in a Zone/subarea were also excluded as were records for divers that had less than eight dives in the whole database. A total of 41 912 data points remained for the analysis. Table 1 gives the number of records used in the final GLM analysis per Model-year and per Zone/subarea.

General Linear Model (GLM) to standardise the CPUE

The following GLM model, which allows for possible annual differences in abalone spatial and temporal distribution, is used to standardise the commercial abalone CPUE data:

$$\ln(CPUE) = \mu + \alpha_{year} + \beta_{season} + \gamma_{zone} + \varphi_{diver} + \eta_{year \times season} + \delta_{year \times zone} + \varepsilon \qquad (1)$$

where:

CPUE	is the catch-per-unit-effort defined as catch (kg) divided by dive time (minutes),
μ	is the intercept,
year	is a factor with 31 levels associated with the Model-years 1980–2011 (excluding 2009 during which the fishery was closed),
season	is a factor with 4 levels associated with the season effect (1 = Jan-Mar; 2 = Apr-Jun; 3 = Jul-Sep; 4 = Oct-Dec),
zone	is a factor with 5 levels associated with the different zones/subareas (A, B, CNP, CP and D),
diver	is a factor with 295 levels associated with the diver code, which includes both the entitlement holders coded in the database as well as "divers". Some recent divers not yet allocated a code were given a temporary code of 555 for the purposes of this analysis[1],
year×*season*	is the interaction between year and season,
year×*zone*	is the interaction between year and zones/subareas, 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 for each zone/subarea is obtained from:

$$CPUE_{year,zone} = \left[\sum_{season}\left(\exp\left(\mu + \alpha_{year} + \beta_{season} + \gamma_{zone} + \varphi_{diver} + \eta_{year \times season} + \delta_{year \times zone}\right)\right)\right] / 4 \qquad (2)$$

where the standardisation is with respect to a fisher code = 8, which contained the most observations as well as the longest period in operation in the fishery.

The reason for standardising in this way when year interactions are present is that the standardised CPUE is to be used as an index of relative abundance when input to assessment models. CPUE itself is assumed to be proportional to local density, so that averaging over season

[1] For the years 2006 to 2011 over which this code was used, such records comprise 0.08% of the total.

is necessary to provide a quantity representative of a consistently calculated average over each year. This averaging is unnecessary in the absence of such interactions, because then the $\exp\left(\alpha_{year}\right)$ term alone will be proportional to abundance.

General Linear Mixed Model (GLMM) to standardise the CPUE

The GLM used to standardise commercial CPUE indices assumes that all factors in the model are fixed effects with the variance of the response values being that of the error term ε. In a GLM analysis we model only the mean (i.e. the fixed effects) of the data. A GLMM has the ability to model not only the mean of the data but also its variance. In fact, a GLMM also allows for the presence of random variables (called random effects) which describe additional variability in the data apart from that reflected by the error term of equation (1). One of the fixed effects factor in the GLM is "divers" with 295 different levels (in the present analysis) associated with different divers with some of the divers in the fishery having very few dives. An alternative approach proposed in this paper is to treat "divers" as a random effect in the GLMM.

The GLMM applied to the abalone commercial CPUE data is of the form:

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

where :

α	is the unknown vector of fixed effects parameters (this vector includes all the parameters for the effects in the GLM model of equation (1) above, except for diver,
X	is the design matrix for the fixed effects,
β	is the unknown vector of random effects parameters (here diver),
Z	is the design matrix for the random effects,
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(β) = E(ε) = 0, so that $E\left(\ln\left(CPUE)\right)\right) = X\alpha$. 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 are 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:

$$\text{Cov}\left(\ln\left(CPUE\right)\right) = \mathbf{V} = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R},$$

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), which produces unbiased estimates for the variance components as it takes into account the degrees of freedom used in estimating the fixed effects. Once estimates of **R** and **G** have been obtained, estimates for the fixed effects parameters (α) can be obtained as well as predictors for the random effects parameters (β).

GLMM standardised CPUE series for each zone/subarea are obtained using the same procedure as that reflected by equation (2).

Results and Discussion

GLM

The examination of the residuals of an initial fit showed evidence of heteroscedasticity, in particular, larger residuals were associated with larger effort (Figure 1). To account for this heteroscedasticity, the iterative, inverse-weighting procedure applied by Plagányi and Edwards (2007) has been applied in which reduced weight is given to the data points with the largest variance in the model.

The GLM model accounts for 46.4% of the total variation of abalone CPUE. Table 2 lists the nominal and standardised CPUE indices provided by the model and Figures 2 shows graphical comparisons of the same. Broadly speaking, the standardisation makes relatively little difference to the nominal trends. Table 3 shows the parameter estimates, together with standard errors, obtained for the single factors included in the GLM model except for the diver factor as there are 295 levels for this factor.

GLMM

Table 4 lists the GLMM-standardised CPUE indices. Figure 3 shows these graphically and compares them to those obtained by a GLM; there is very little difference between the results for the two different methods of standardisation.

Reference

Brandão, A. and Butterworth, D.S. 2009. A summary of the General Linear Model analyses applied to the commercial abalone CPUE data for Zones A-D over the period 1980-2008. Marine and Coastal Management document: MCM/2009/OCT/SWG-AB/06.

Plagányi, É. and Edwards, C. 2007. Summary of the GLM used to standardise abalone catch-per-unit-effort data for Zones A-D over the period 1980-2006. Marine and Coastal Management document: WG/AB/07/Aug/19.

Table 1. The number of data entries per Zone available for the final GLM analysis to standardise the commercial abalone CPUE series are shown. Subarea CNP was closed during the 2001 fishing season and subarea CP during both the 2001, 2002 and 2003 fishing seasons. The abalone fishery was closed in February 2008 and reopened in 2010. Some sample sizes were considered too small and were not included in the analysis (see text). Model-years are defined as the period October of the preceding year to September of the year indicated.

Model year	**Zone/subarea**				
	A	**B**	**CNP**	**CP**	**D**
1980	257	555	73	753	535
1981	192	578	147	622	383
1982	311	610	109	594	608
1983	327	690	144	466	301
1984	334	696	274	364	373
1985	359	619	158	366	583
1986	340	763	222	445	205
1987	443	585	105	494	144
1988	457	434	96	498	147
1989	447	414	91	504	184
1990	525	410	138	458	140
1991	446	403	161	539	167
1992	347	302	98	396	142
1993	299	237	110	333	75
1994	345	290	155	287	162
1995	441	238	137	333	170
1996	506	324	401	427	206
1997	720	248	249	116	194
1998	599	472	205	71	290
1999	667	412	56	8	295
2000	447	319	23		302
2001	387	286			133
2002	285	226	96		94
2003	413	128	54		25
2004	94	567	155		60
2005	61	596	167		52
2006	41	671	164		48
2007		482			
2008		291			
2009					
2010	175	229			
2011	335	357			

Table 2. Nominal and GLM-standardised commercial CPUE series for abalone for Model-years (October of the preceding year to September of the year indicated) 1980 to 2011 and Zones/subareas A, B, CNP, CP and D. Both the nominal and the standardised values have been divided by the mean value of the respective series.

a) Nominal CPUE series

Model year	Zone/subarea				
	A	B	CNP	CP	D
1980	1.060	0.785	0.873	0.840	0.907
1981	1.034	0.788	0.897	0.834	0.833
1982	0.908	0.797	0.881	0.834	0.803
1983	0.895	0.777	0.939	0.875	0.723
1984	0.972	0.830	0.962	0.891	0.796
1985	0.912	0.841	0.919	0.965	0.809
1986	1.010	0.916	1.023	1.103	0.770
1987	1.039	0.899	1.144	1.056	0.868
1988	1.115	0.986	1.185	1.148	1.033
1989	1.020	1.002	1.154	1.116	0.894
1990	1.169	1.222	1.417	1.215	1.245
1991	1.169	1.266	1.222	1.102	1.231
1992	1.292	1.329	1.263	1.234	1.163
1993	1.397	1.579	1.089	1.295	1.909
1994	1.327	1.365	1.219	1.324	1.709
1995	1.234	1.472	1.252	1.130	1.476
1996	1.225	1.391	0.979	0.904	1.439
1997	1.134	1.479	0.892	0.727	1.495
1998	1.172	1.328	0.983	0.737	1.548
1999	1.011	1.168	0.989	0.672	1.035
2000	1.044	1.180	1.092		0.950
2001	1.037	1.089			0.848
2002	1.035	1.117	1.278		0.769
2003	0.842	1.061	0.802		0.489
2004	0.817	0.781	0.568		0.449
2005	0.512	0.701	0.519		0.391
2006	0.544	0.591	0.458		0.418
2007		0.510			
2008		0.503			
2009					
2010	0.628	0.753			
2011	0.447	0.493			

b) GLM-standardised CPUE series

Model year	Zone/subarea				
	A	B	CNP	CP	D
1980	1.094	0.867	0.960	0.946	0.962
1981	1.063	0.850	1.019	0.925	0.879
1982	0.930	0.873	1.015	0.912	0.855
1983	0.902	0.823	0.985	0.929	0.753
1984	0.972	0.854	0.985	0.942	0.820
1985	0.917	0.863	0.944	0.967	0.832
1986	0.947	0.943	1.056	1.136	0.899
1987	0.986	0.877	1.108	1.007	0.949
1988	1.021	0.955	1.115	1.069	1.063
1989	1.038	1.019	1.157	1.121	0.995
1990	1.077	1.118	1.271	1.127	1.209
1991	1.091	1.063	1.161	1.062	1.044
1992	1.241	1.283	1.267	1.219	1.197
1993	1.153	1.393	1.080	1.295	1.653
1994	1.118	1.065	1.194	1.205	1.299
1995	1.157	1.231	1.187	1.120	1.206
1996	1.216	1.325	1.050	0.983	1.317
1997	1.162	1.478	0.846	0.722	1.418
1998	1.137	1.343	0.960	0.694	1.461
1999	1.060	1.289	1.053	0.619	1.084
2000	1.043	1.163	1.059		0.955
2001	1.099	1.153			0.879
2002	1.142	1.140	1.208		0.806
2003	0.922	1.062	0.749		0.557
2004	0.916	0.895	0.578		0.700
2005	0.672	0.776	0.527		0.557
2006	0.758	0.713	0.464		0.650
2007		0.643			
2008		0.566			
2009					
2010	0.652	0.796			
2011	0.514	0.581			

Table 3. Parameters estimates and standard errors for the single factors *Year*, *Season* and *Zone* included in the GLM to obtain standardised indices of abundance for abalone.

	Parameter estimate	**Standard error**
Year		
1980	0.000	—
1981	0.032	0.024
1982	0.041	0.025
1983	0.055	0.025
1984	0.062	0.025
1985	0.103	0.026
1986	0.081	0.025
1987	0.076	0.027
1988	0.162	0.030
1989	0.168	0.029
1990	0.400	0.029
1991	0.319	0.032
1992	0.497	0.039
1993	0.527	0.048
1994	0.384	0.040
1995	0.330	0.041
1996	0.540	0.030
1997	0.533	0.044
1998	0.586	0.026
1999	0.486	0.029
2000	0.455	0.032
2001	0.427	0.033
2002	0.388	0.034
2003	0.321	0.045
2004	0.322	0.044
2005	0.122	0.029
2006	-0.069	0.029
2007	-0.185	0.046
2008	-0.486	0.035
2009	—	—
2010	-0.019	0.035
2011	-0.394	0.032
Season		
Jan-Mar	0.000	—
Apr-Jun	-0.021	0.020
Jul-Sep	0.116	0.020
Oct-Nov	0.157	0.043
Zone		
A	0.386	0.028
B	0.000	—
CNP	-0.056	0.048
CP	-0.059	0.022
D	0.146	0.023

Table 4. GLMM-standardised commercial CPUE series for abalone for Model-years (October to September) 1980 to 2011 and Zones/subareas A, B, CNP, CP and D. The GLMM standardised values have been divided by the mean value of the respective series.

Model year	Zone/subarea				
	A	B	CNP	CP	D
1980	1.070	0.839	0.941	0.914	0.946
1981	1.069	0.849	1.009	0.924	0.881
1982	0.930	0.873	1.014	0.913	0.863
1983	0.906	0.826	0.985	0.928	0.752
1984	0.968	0.859	0.977	0.938	0.827
1985	0.919	0.871	0.932	0.977	0.842
1986	0.944	0.941	1.038	1.122	0.904
1987	0.992	0.892	1.108	1.025	0.962
1988	1.011	0.956	1.106	1.080	1.076
1989	1.038	1.022	1.133	1.127	1.008
1990	1.084	1.126	1.269	1.136	1.226
1991	1.100	1.064	1.146	1.074	1.056
1992	1.251	1.290	1.275	1.231	1.213
1993	1.175	1.407	1.090	1.276	1.672
1994	1.106	1.051	1.184	1.204	1.266
1995	1.186	1.245	1.201	1.118	1.217
1996	1.235	1.354	1.055	1.002	1.343
1997	1.169	1.481	0.863	0.728	1.430
1998	1.157	1.357	0.968	0.694	1.476
1999	1.061	1.287	1.020	0.590	1.084
2000	1.045	1.154	1.058		0.959
2001	1.065	1.115			0.862
2002	1.114	1.108	1.148		0.797
2003	0.918	1.051	0.781		0.561
2004	0.908	0.875	0.638		0.664
2005	0.640	0.748	0.566		0.505
2006	0.720	0.698	0.494		0.607
2007		0.628			
2008		0.586			
2009					
2010	0.715	0.875			
2011	0.501	0.571			

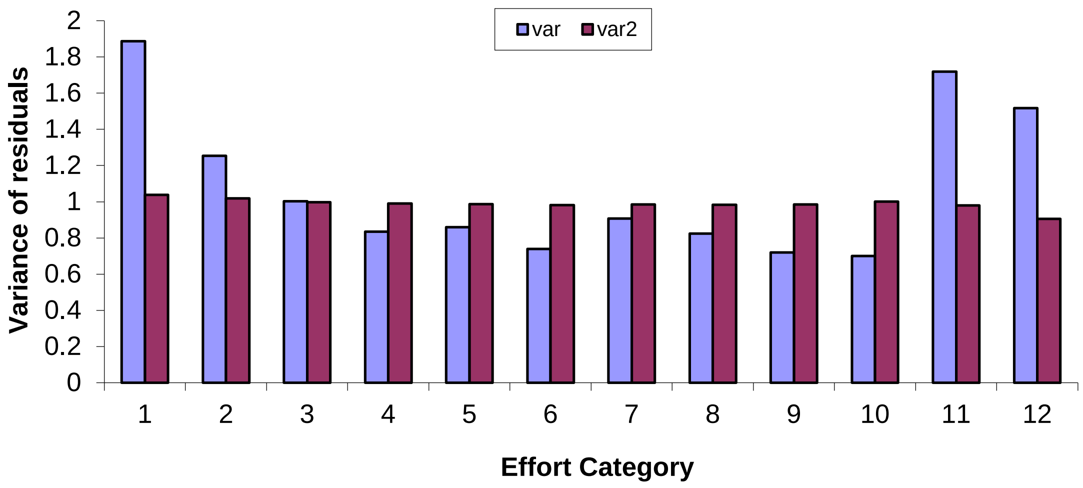

Figure 1. Plot of the average variance of the residuals versus effort category (50 minutes interval) for an initial GLM fit of the data (var, showing evidence of heteroscedasticity) and for the final weighted GLM (var2, showing homoscedastic residuals).

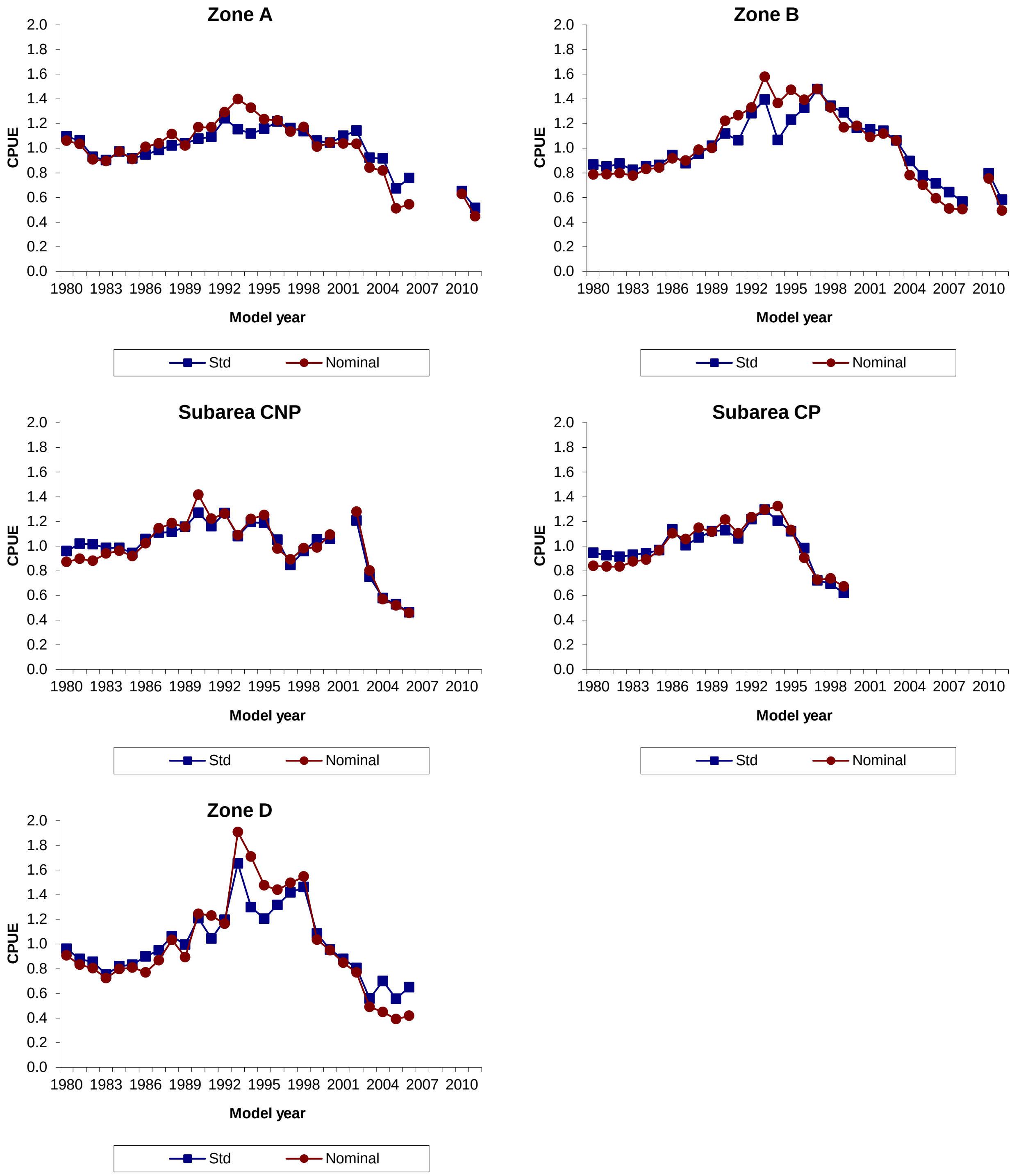

Figure 2. GLM-standardised CPUE trends (normalised to their means over the 31 year period) for Zones/subareas A, B, CNP, CP and D. For comparison, the nominal series (also normalised to their means over the 31 year period) is also shown.

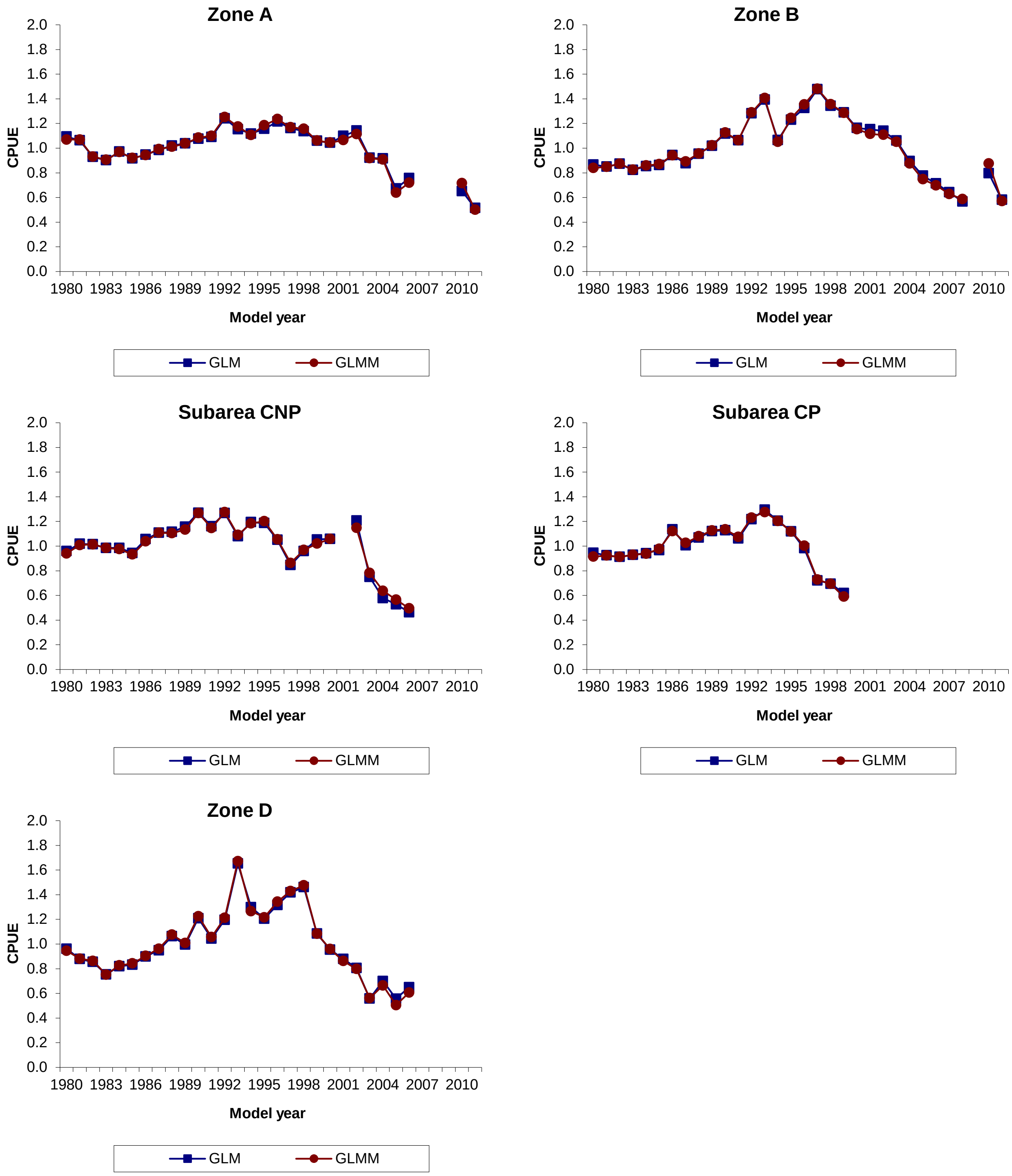

Figure 3. GLMM-standardised CPUE trends (normalised to their means over the 31 year period) for Zones/subareas A, B, CNP, CP and D. For comparison, the GLM-standardised series (also normalised to their means over the 31 year period) is also shown.