The application of a Generalized Linear Mixed Model to the Area 8+ trapboat data

J.P. Glazer and D. S. Butterworth

Introduction

Two papers by Glazer and Butterworth (2011a, 2011b) were recently submitted to the West Coast Rock Lobster Scientific Working Group meeting for consideration. The first of those included sub-area as an explanatory variable in a Generalized Linear Mixed Model (GLMM) applied to standardize the Area 8 trapboat CPUE data. The second paper incorporated data from Area 8+ (rather than data from Area 8 only as used in the past) in the standard Generalized Linear Model (GLM) applied to those data. It was agreed at the meeting that the analyses presented in the aforementioned papers be amalgamated by conducting a GLMM on the Area 8+ data, which incorporates sub-area as an explanatory variable in the model. The results of such an analysis are presented in this paper.

It should be noted that Area 8+ for trapboat fishing comprises data from Areas 8, 10 and 11. No trapboat fishing takes place in the area East of Hangklip (Areas 12-14) and a procedure has therefore been adopted to adjust the standardized index to allow for the movement of lobster into the East of Hangklip area over a period of time (1987-1995). The index obtained above is then extended back to 1985 by scaling the pre-1992 indices from the model applied in the past to standardize the Area 8 CPUE data, so that they can be incorporated with the GLMM-based index.

The data

Catch and effort data are available since 1981 for Area 8 and since 1992 for Areas 10 and 11 respectively. In addition, information at a sub-area level is available since 1992 only. The GLMM analyses reported here are therefore restricted to data since 1992.

Certain general data exclusions have been applied prior to the application of the GLMM. These are as follows:

- Month=October (historically very little fishing took place in this month)
- Hout Bay Fishing vessels that fished over the period 1997-2000 (catch data incorrect)
- Effort (traps)=0
- Catch=0

The sample sizes per year and month are shown in Table 1. In past analyses of Area 8 data from all months were included in the analyses, but the resulting standardized index was calculated from a core set of those months (January to June), particularly given the patchiness of data in the non-core months. It was also shown in Glazer and Butterworth (2011a) that trends obtained from an analysis of the data for all months *vs* core months were identical, and further analyses reported in that paper were restricted to data from the core months only. Based on the sample sizes shown in Table 1, it would seem reasonable to include data from the months January – July (August was omitted because of considerations related to modelling random effects, as explained below). Table 2 therefore shows the sample sizes per year and sub-area for that period.

It is common practice in local GLMM analyses of CPUE data to exclude data from cells that have few samples taken in them (usually $n \leq 5$), given that the inclusion of these data may lead to anomalous results if the few

samples available are atypical; hence the shaded cells in Table 2 have been omitted from the analyses. Furthermore, sub-area 3 of Area 8 has also been omitted from the analyses due to the patchiness of data over time in that sub-area.

The GLMM and associated results

A model of the form shown in equation (1) was applied to the Area 8+ data from 1992 onwards.

$\ell n(\text{CPUE})=\alpha+\beta_{year}+\gamma_{\text{month}}+\eta_{sub\text{-}area}+(\text{year} \times \text{month})+(\text{year x subarea})+ \varepsilon$ (1)

α is the intercept,
year is a factor with 18 levels (1992-2009) associated with the year effect,
month is a factor with 7 levels (January-July) associated with the month effect, and
sub-area is a factor with 7 levels (subareas 1-2, 4-6 in Area 8, Area 10, Area 11) associated with the sub-area effect.

Both the month and sub-area interactions with year are treated as random effects.

In order to derive an index of abundance the model is run twice; the second run excluding records where the residuals from the first run exceed ± 2SD. This methodology was adopted in order to adjust for outliers (leading to non-normality of the residuals) evident in the initial model run.

The exponent of the year factors, adjusted for movement of lobster into the East of Hangklip area, is taken to be the standardized CPUE index, i.e. $CPUE_{year} = e^{year} \times (\frac{A_{8+,year}}{A_{8+}})$. The proportion $(\frac{A_{8+,year}}{A_{8+}})$ is applied to adjust the Area 8+ area size (3927.31km^2) to include East of Hangklip (comprising a total area size of 161.96km^2). $A_{8+,year}$ is year-specific (the Area 8+ size is expanded in a linear fashion over the period 1987-1995) and A_{8+} is the area size of Area 8+. The resultant year-specific proportions applied to the exponent of the year factors are as follows:

Year	proportion
≤1986	1
1978	1.005
1988	1.009
1989	1.014
1990	1.018
1991	1.023
1992	1.027
1993	1.032
1994	1.037
≥1995	1.041

The standardized index, together with the nominal trend, is shown in Figure 1.

The GLM-standardized index used in the past ("Revised Area 8" in Figure 1 of Glazer and Butterworth, 2011b) incorporates data from 1985 given that the model does not include a sub-area effect. A method of combining the GLMM index with that of the GLM index was considered desirable in order to extend the series as far back in time as possible. This was achieved by multiplying the pre-1992 GLM values by the ratio $\frac{\overline{Std\ CPUE_{GLMM,1992-1996}}}{\overline{Std\ CPUE\ _{GLM,1992-1996}}}$ in order to scale them to the GLMM index and then combine them with the GLMM index. The resulting combined index is reported in Table 3 and shown in Figure 2.

The GLMM fitted assumes that the random effects are homoscedastic and uncorrelated. Figures 3 and 4 show the random effects by month and by sub-area respectively. There is no obvious indication of substantial non-randomness. The random effects for August did however show a systematic pattern – hence those data were omitted from the analysis.

The assumption of normality of the error term was investigated by examining the unstandardized residuals obtained from the GLMM fit after the exclusion of outliers. The mean, median and mode are 0, 0.03 and 0.1 respectively. The skewness and kurtosis statistics (which for a normal distribution should equal 0) are -0.25 and -0.26 respectively. Given that the median (0.03) is much less than the standard deviation of the residuals (0.52), the non-normality of the residual distribution is probably not too much of a cause for concern. The residual distribution is shown in Figure 5.

References

Glazer, J.P and D. S. Butterworth. 2011a. The inclusion of sub-areas in the standardization of the Area 8 trapboat CPUE data through the application of a Generalized Linear Mixed Model. Unpublished Working Group Document: Fisheries/2011/MAR/SWG-WCRL 03. 11pp.

Glazer, J.P and D. S. Butterworth. 2011b. Updated GLM analyses of Area 8+. Unpublished Working Group Document: Fisheries/2011/MAR/SWG-WCRL 04. 10pp.

Table 1: Sample sizes per year and month for Areas 8, 10 and 11. Data from the shaded cells are included in the GLMM analyses of Area 8+

	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Total
1992	4	47	113	212	208	249	297	181	62	61		1434
1993	4	94	22	122	176	213	299	140	290	145		1505
1994	4	51	279	249	190	313	237	138	72	38	13	1584
1995	5	22	49	171	288	184	236	186	148	54		1343
1996		5	138	223	225	215	198	244	432	109	7	1796
1997			43	61	215	190	413	337	253	149	54	1715
1998		18	28	36	164	175	171	333	413	247	248	1833
1999		8	22	121	174	386	360	242	172	146	90	1721
2000		1	9	24	143	165	393	285	207	110	125	1462
2001		2	10	29	175	234	181	236	342	571	621	2401
2002	4	24	33	53	78	159	232	242	359	364	608	2156
2003	7	12	48	154	318	309	349	311	383	391	306	2588
2004	19	25	20	84	214	310	344	466	426	500	670	3078
2005					90	311	203	793	390	270	342	2399
2006	17	42	56	75	476	380	708	294	421	769	818	4056
2007	1	18	164	162	244	381	183	646	330	511	453	3093
2008		18	147	90	257	323	352	349	531	259	301	2627
2009		26	153	231	521	332	267	199				1729
Total	65	413	1334	2097	4156	4829	5423	5622	5231	4694	4656	38520

Table 2: Sample sizes per year and sub-area for the January to July period. Data from cells where n ≤ 5 (shaded areas) are omitted from the analyses.

	SA1	SA2	SA3	SA4	SA5	SA6	A10	A11	Total
1992	248	590		233	41	73	76	61	1322
1993	363	413	18	302	68	18	15	65	1262
1994	523	546	13	211	66	22	54	43	1478
1995	628	357		80	28	11	109	49	1262
1996	601	447	38	296	33	45	133	82	1675
1997	534	613	22	98	41	71	131	2	1512
1998	243	736	43	133	14	22	114	15	1320
1999	347	580	46	267	5	47	152	33	1477
2000	560	215	62	188	24	45	121	11	1226
2001	602	366	17	91	1	9	105	16	1207
2002	491	269		222	18	41	77	38	1156
2003	757	480		265	141	95	86	48	1872
2004	663	336		256	61	397	76	75	1864
2005	124	418		414	95	536	124	76	1787
2006	172	313		699	34	954	164	74	2410
2007	260	436		564	133	391	196	130	2110
2008	141	342	1	675	189	361	222	118	2049
2009	216	626	1	412	7	75	297	69	1703
Total	7473	8083	261	5406	999	3213	2252	1005	28692

Table 3: Trapboat standardized CPUE indices for Area 8+.

Year	CPUE
1985	0.955
1986	1.316
1987	1.102
1988	1.232
1989	1.044
1990	0.478
1991	0.809
1992	1.023
1993	1.257
1994	1.101
1995	1.498
1996	1.152
1997	1.502
1998	1.363
1999	1.497
2000	1.582
2001	1.844
2002	2.171
2003	1.655
2004	1.562
2005	1.177
2006	1.029
2007	0.926
2008	1.042
2009	1.041

Figure 1: Area 8+ standardized CPUE index. The nominal CPUE trend is also shown. Both indices have been normalized to their respective means.

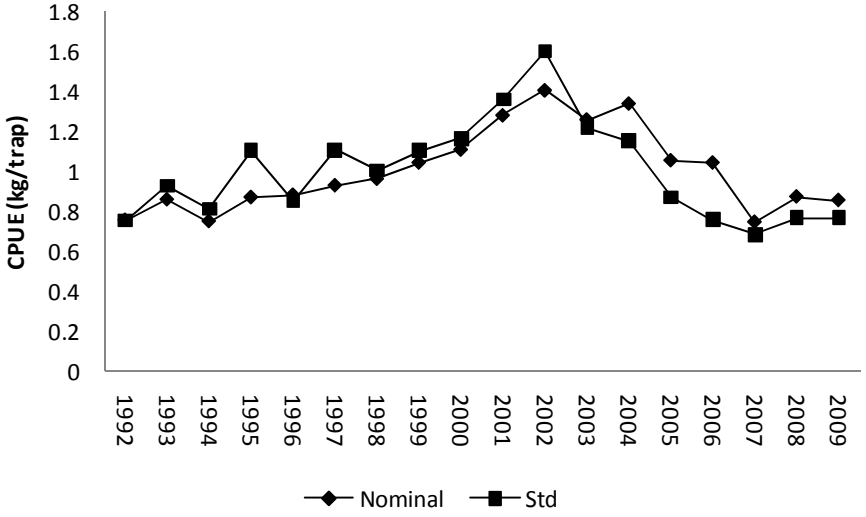

Figure 2: Area 8+ standardized CPUE index extended back to 1985.

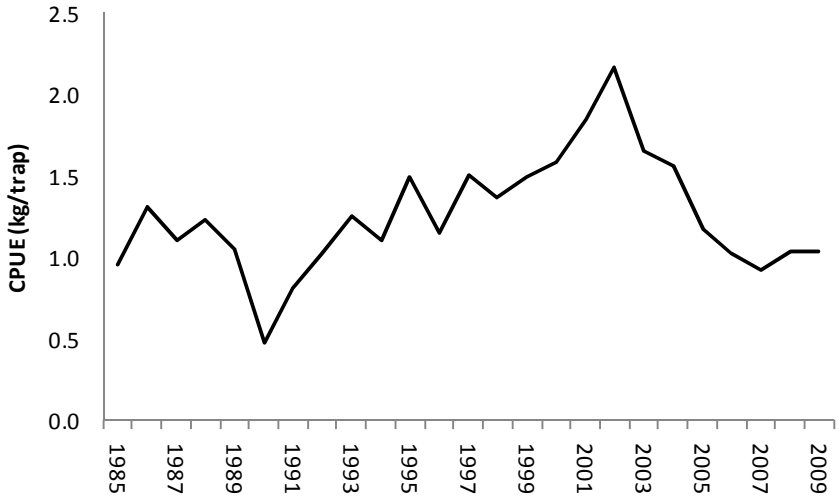

Figure 3: Random effect estimates by month obtained from the GLMM.

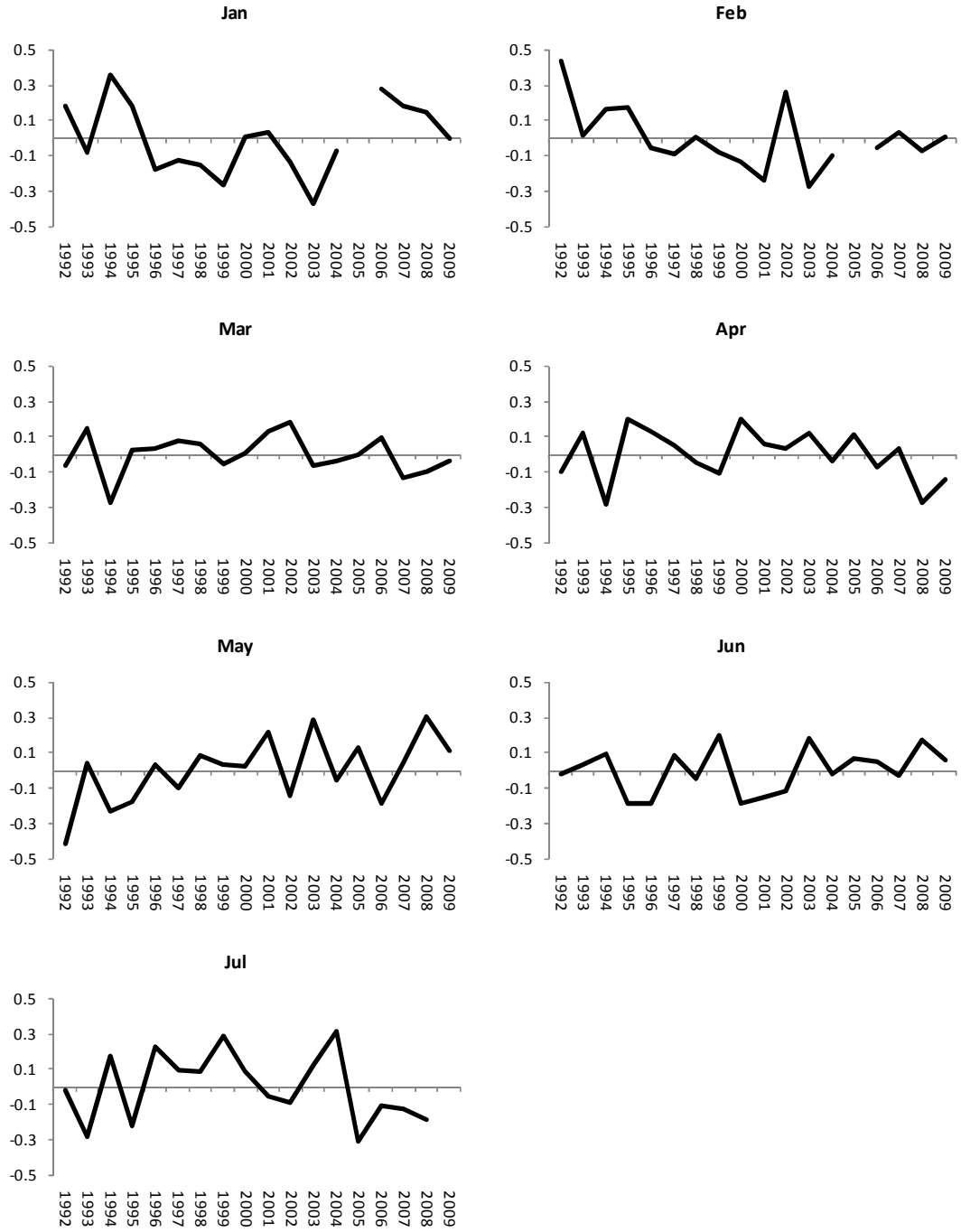

Figure 4: Random effect estimates by sub-area obtained from the GLMM.

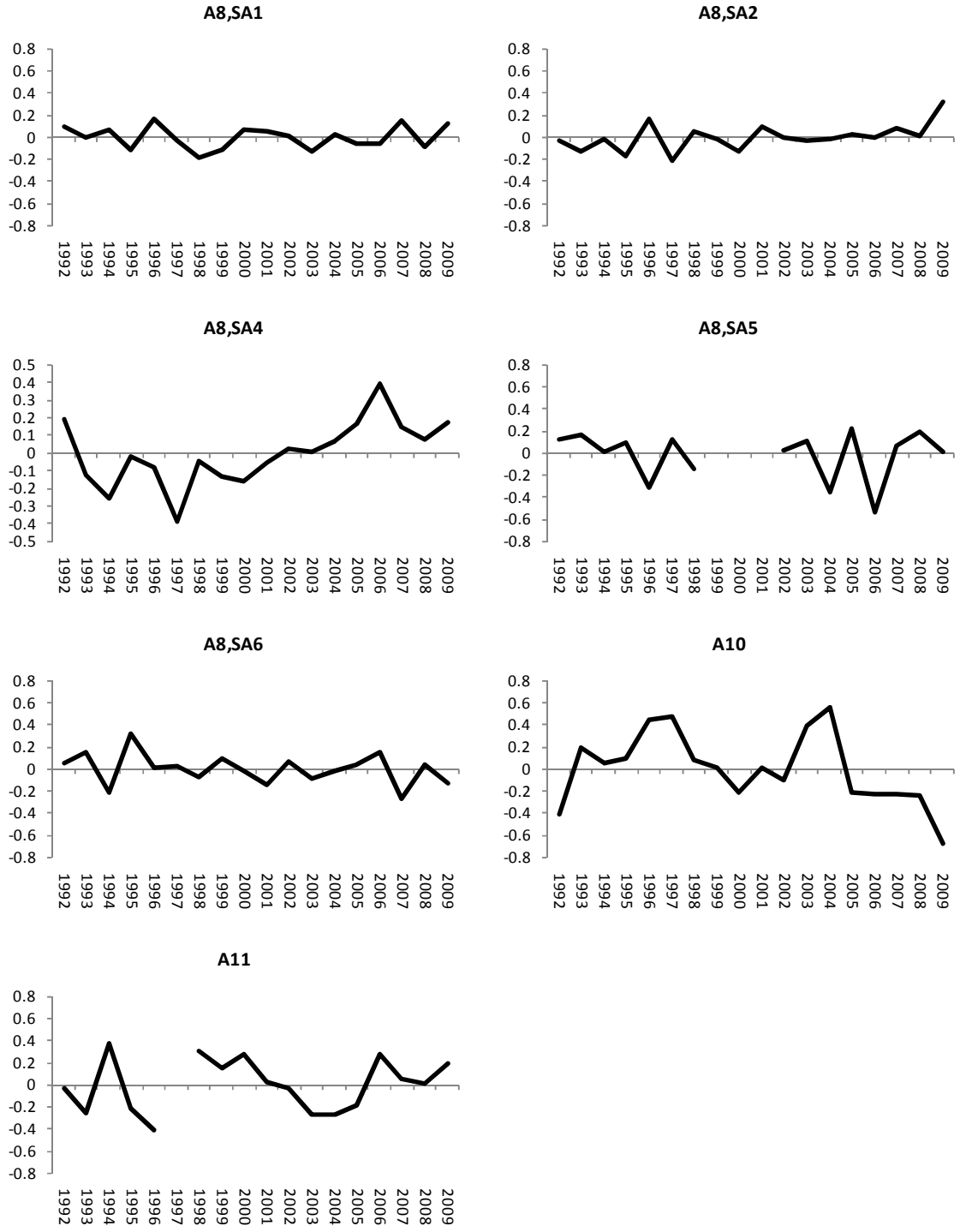

Figure 5: Distribution of unstandardized residuals obtained from the GLMM.