Standardisation of the commercial CPUE series for Abalone in Zones E and G from 1980 to 2007 using a mixed effects model

Charles Edwards, Éva Plagányi,
Anabela Brandão and Doug Butterworth
Marine Resource Assessment and Management group,
Department of Mathematics and Applied Mathematics,
University of Cape Town
August, 2007

Introduction

The commercial CPUE series is considered during modeling of resource dynamics as an index of population abundance. However, a number of factors other than abundance may influence recorded values. Standardisation is able to take into account some of these effects, thereby producing a more reliable index.

Methods

Commercial Catch per Unit Effort (CPUE) data (including Limited Divers landings) from 1980 to 2007 was supplied by Angus Mackenzie (Marine and Coastal Management). Additional information that could potentially by used during standardisation included the area, date and diver number for each CPUE record. The date was considered as a discrete factor in terms of the model year (running from October of the previous year until September of the current year) and four three month seasons.

A total of 1031 CPUE records were available for Zone E and 1431 for Zone G. The data was first cleaned of likely errors by plotting the number of abalone landed against the recorded catch in kilograms (on a log scale), and removing outliers. In this way two likely errors were removed from the Zone E data but none from Zone G. An additional one record with no date from Zone E, five records from Zone G with no diver number and two records from Zone G with zero CPUE values were excluded from the analysis. A total of 1028 data points were therefore included in the analysis for Zone E and 1424 for Zone G.

Including random effects

In previous standardisations of the CPUE series for Abalone we have sought to estimate the effect sizes for a range of factors, such as year, area and season. These factors contribute to variation in the CPUE. By accounting for this variation we are able to ensure that they do not influence our estimation of the CPUE for a particular year. Standardising in this way therefore makes the CPUE trend across years a more reliable index of population abundance.

In the standardisation presented here, we treat Diver as a random effect. This means that the effect of Diver on $lnCPUE$ is considered to be a (normally distributed) random variable with variance σ^2_D. Discrete factors included in the model are considered to have a multiplicative effect on CPUE. We have therefore used the natural logarithm of the CPUE ($lnCPUE$) during standardisation. The mixed effects model is represented as:

$$lnCPUE_{ij} = \mu + D_i + \alpha + \cdots + \gamma + \epsilon_{ij}$$

where,

$lnCPUE_{ij}$ is the j^{th} observation for the i^{th} Diver;

μ is the average across all factors;

D_i is a continuous random variable with $D_i \sim N(0, \sigma^2_D)$;

$\alpha \ldots \gamma$ are fixed effects included in the model; and,

ϵ_{ij} is the residual error with $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$.

This can be expressed in matrix notation for n observations taken from q divers with a combined $p-1$ levels for all fixed effects:

$$\mathbf{lnCPUE} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZD} + \boldsymbol{\epsilon} \tag{1}$$

where,

$\boldsymbol{\beta}$ is a length p vector containing the intercept plus fixed effect coefficients;

$\mathbf{X}$ is a $n \times p$ matrix defining the contribution of each coefficient to **lnCPUE**;

$\mathbf{D}$ is a length q vector of random effects;

$\mathbf{Z}$ is a $n \times q$ matrix relating each random effect to different divers; and,

$\boldsymbol{\epsilon}$ is the residual error with $\boldsymbol{\epsilon} \sim N(0, \sigma^2_\epsilon \mathbf{I_{n \times n}})$.

Observed **lnCPUE** values are distributed as:

$$\mathbf{lnCPUE} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{2}$$

where,

$\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix.

The covariance matrix is estimated as:

$$\boldsymbol{\Sigma} = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}' + \sigma_\epsilon^2\mathbf{I_{n\times n}} \tag{3}$$

where,

$\boldsymbol{\Gamma}$ is a diagonal $q \times q$ matrix describing the variance due to the Diver random effect with $\boldsymbol{\Gamma} = \sigma_D^2\mathbf{I_{q\times q}}$,

such that,

$\Sigma_{uv} = \sigma_\epsilon^2 + \sigma_D^2$ for $u = v$;

$\Sigma_{uv} = \sigma_D^2$ for when $u \neq v$ but observations u and v are from the same diver;

$\Sigma_{uv} = 0$ otherwise.

During maximum likelihood estimation we minimse the log-likelihood of the error:

$$lnL(lnCPUE) = ln|\boldsymbol{\Sigma}| + \mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{e} \tag{4}$$

where,

$\mathbf{e}$ is the vector of errors attributable to the fixed effects: $\mathbf{e} = (\mathbf{lnCPUE} - \mathbf{X}\boldsymbol{\beta})$.

However because Maximum Likelihood can underestimate the variance attributable to the fixed effects ($\sigma_\epsilon^2\mathbf{I_{n\times n}}$) we generally use Restricted Maximum Likelihood estimation, minimising:

$$lnL(lnCPUE) = ln|\boldsymbol{\Sigma}| + \mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{e} + ln|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}|. \tag{5}$$

Incorporating random effects in this way substantially reduces the number of parameters to be estimated, thereby improving statistical power.

Results

We first test whether introducing Diver as a random effect leads to a significant improvement in model fit. This involved fitting two models for each Zone using Restricted Maximum Likelihood estimation and comparing their explanatory power with the AIC.

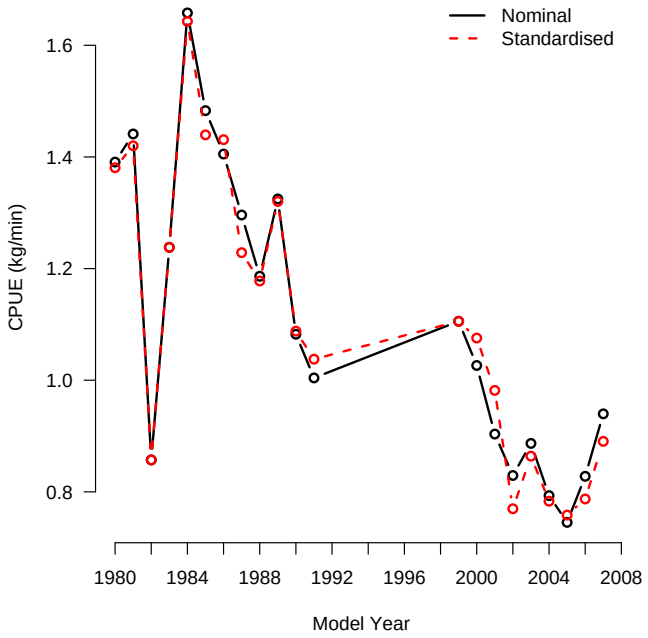

Figure 1: Nominal and standardised CPUE series plotted against Model Year: **Zone E**.

Model 1 $lnCPUE = \mu + D + \alpha_{YEAR} + \beta_{AREA} + \delta_{SEASON} + \varepsilon$

Model 2 $lnCPUE = \mu + \alpha_{YEAR} + \beta_{AREA} + \delta_{SEASON} + \varepsilon$

The AIC is estimated as $AIC = -2lnL+2k$ where k is the number of parameters estimated. When considering a random effects model such as Model 1, we are estimating one extra parameter, namely σ^2_D. For Zone E Model 1, $AIC = 363.61$ and for Zone E Model 2, $AIC = 519.68$, indicating that a significant amount of the variation in $lnCPUE$ can be explained by variation between divers. For Zone G Model 1, $AIC = 615.87$ and for Zone G Model 2, $AIC = 814.21$. We therefore retained the mixed effects model (Model 1). An initial examination of the distribution of standardised residuals identified a single outlier for Zone E, two for Zone G and an additional influential observation for Zone G that were excluded from further analysis.

We next examined significance of the factors α_{YEAR}, β_{AREA} and δ_{SEASON}. This involved fitting nested models using Maximum Likelihood estimation and comparing with the AIC.

Model 1 $lnCPUE = \mu + D + \alpha_{YEAR} + \beta_{AREA} + \delta_{SEASON} + \varepsilon$

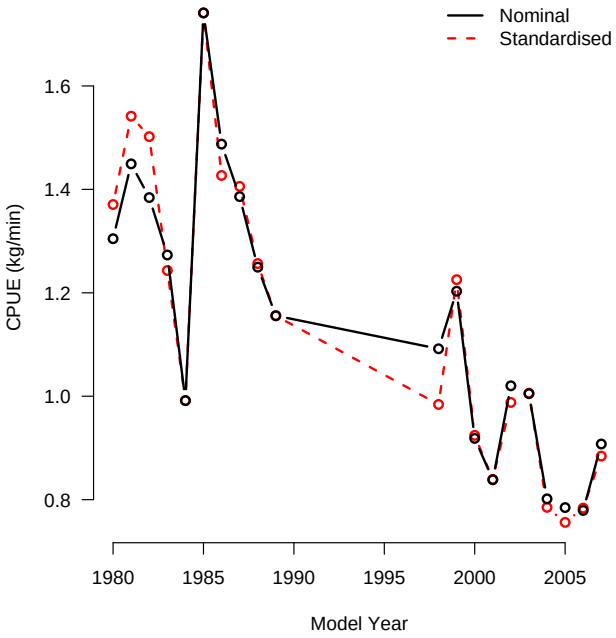

Figure 2: Nominal and standardised CPUE series plotted against Model Year: **Zone G**.

Model 2 $lnCPUE = \mu + D + \alpha_{YEAR} + \beta_{AREA} + \varepsilon$

Model 3 $lnCPUE = \mu + D + \alpha_{YEAR} + \varepsilon$

We obtained the following AIC values for the different model fits:

	Zone E AIC	Zone G AIC
Model 1	-17.97	195.25
Model 2	-22.21	190.67
Model 3	9.70	197.41

Model 2 has the lowest AIC for both Zones, and we therefore adopted it as the best representation of the data and re-fitted using Restricted Maximum Likelihood (giving $AIC = 299.69$ for Zone E and $AIC = 490.57$ for Zone G). Testing assumptions made by the model we found that the Diver random effects did not differ significantly from normality for either Zone (Shapiro-Wilk normality test). However examination of the distributions of standardised residuals revealed heterodescasity with increased variances at lower $lnCPUE$ values.

Accounting for heterodescasity

If we consider fishing to be a Poisson process, so that Abalone are encountered at random and at a constant rate per unit effort during a particular dive, then we expect variance to be inversely related to dive time. Plotting the absolute residuals against dive time revealed this to be approximately true. This trend in variance can be compensated for by grouping dives into different effort categories and estimating the variance in each group. These variances are used to weight the contributions of different data points to the log-likelihood, with $w_{ic} = 1/\sigma_c^2$ for observation i and effort category c. The covariance matrix $\mathbf{\Sigma}$ is therefore described by σ_D^2, σ_ϵ^2 and a length n diagonal vector of weights $\mathbf{w}$, which are all coestimated during the fitting process.

Zone E

Assuming 10 effort categories and coestimating the weights for each category led to an improved model fit with $AIC = 283.59$. However further examination of the residuals revealed a positive relationship between variance and Model Year. We therefore repeated the weighting procedure instead estimating weights for each Model Year. This resulted in a substantially improved model fit, with $AIC = 223.66$, and homodescastic standardised residuals. Estimating weights for every combination of effort and year categories merits attention, but due to its likely small effect we here adopted the model weighted by year to provide a final standardisation of the CPUE series for Zone E.

Zone G

Coestimating the weights for each of 10 effort categories led to an improved model fit with $AIC = 466.07$. Although a positive relationship between between variance and Model Year was observed, convergence was not reached during estimation of the weights for each year. We therefore used the model weighted according to effort category to provide a final standardisation of the CPUE series for this zone, although slight heterodescasticity remained.

The standardised CPUE series for Zone E is listed in Table 1 and illustrated in Figure 1. Estimated coefficients are given in Table 3. Note that coefficients have been estimated relative to the overall mean μ (rather than to the mean of a particular level). The reported p values therefore provide a reliable indication of significance. Standard errors where estimated as $\sigma_D = 0.126$ and $\sigma_\epsilon = 0.175$. A

reasonable proportion of variation was explained by the model with $R^2 = 0.67$. For Zone G, the standardised CPUE series is listed in Table 2 and illustrated in Figure 2. Estimated coefficients are given in Table 4. Standard errors where estimated as $\sigma_D = 0.154$ and $\sigma_\epsilon = 0.232$, with $R^2 = 0.58$.

Conclusion

Standardisation of the commercial CPUE series provides a more reliable index of population abundance. Being the primary input into the stock assessment models used in Zones E and G makes this standardisation particularly important. The results presented here were used in subsequent modeling of resource dynamics for the 2007 Model Year (WG/AB/Aug/25).

Table 1: Standardised commercial CPUE series: **Zone E**.

Model Year	n	Nominal	Standardised
1980	19	1.39	1.38
1981	8	1.44	1.42
1982	2	0.86	0.86
1983	1	1.24	1.24
1984	8	1.66	1.64
1985	160	1.48	1.44
1986	9	1.41	1.43
1987	43	1.30	1.23
1988	16	1.19	1.18
1989	42	1.32	1.32
1990	19	1.08	1.09
1991	42	1.00	1.04
1992			
1993			
1994			
1995			
1996			
1997			
1998			
1999	25	1.11	1.11
2000	32	1.03	1.08
2001	28	0.90	0.98
2002	73	0.83	0.77
2003	43	0.89	0.86
2004	141	0.79	0.78
2005	132	0.75	0.76
2006	114	0.83	0.79
2007	70	0.94	0.89

Table 2: Standardised commercial CPUE series: **Zone G**.

Model Year	n	Nominal	Standardised
1980	9	1.30	1.37
1981	11	1.45	1.54
1982	18	1.38	1.50
1983	9	1.27	1.24
1984	1	0.99	0.99
1985	1	1.74	1.74
1986	89	1.49	1.43
1987	76	1.39	1.41
1988	95	1.25	1.26
1989	99	1.16	1.16
1990			
1991			
1992			
1993			
1994			
1995			
1996			
1997			
1998	91	1.09	0.98
1999	17	1.20	1.23
2000	39	0.92	0.92
2001	98	0.84	0.84
2002	109	1.02	0.99
2003	118	1.00	1.01
2004	152	0.80	0.79
2005	175	0.78	0.76
2006	155	0.78	0.78
2007	59	0.91	0.88

Appendix

Table 3: Estimated coefficients : **Zone E**.

Coefficient	Estimate	Std. error	df	t-value	$Pr(>\|t\|)$
Intercept	0.056	0.034	876	1.665	0.09640
year1	0.014	0.030	876	0.467	0.64070
year2	-0.082	0.061	876	-1.329	0.18420
year3	0.122	0.059	876	2.078	0.03800
year4	0.033	0.018	876	1.814	0.07010
year5	0.000	0.011	876	0.025	0.97970
year6	-0.001	0.011	876	-0.057	0.95490
year7	-0.021	0.007	876	-2.851	0.00450
year8	-0.016	0.007	876	-2.252	0.02460
year9	-0.008	0.005	876	-1.710	0.08760
year10	-0.024	0.007	876	-3.550	0.00040
year11	-0.019	0.004	876	-5.246	<0.00001
year12	-0.014	0.004	876	-3.778	0.00020
year13	-0.020	0.004	876	-5.789	<0.00001
year14	-0.027	0.004	876	-6.528	<0.00001
year15	-0.027	0.003	876	-8.379	<0.00001
year16	-0.021	0.003	876	-6.567	<0.00001
year17	-0.020	0.002	876	-9.945	<0.00001
year18	-0.019	0.002	876	-10.605	<0.00001
year19	-0.015	0.002	876	-8.553	<0.00001
year20	-0.008	0.002	876	-4.361	<0.00001
area1	-0.313	0.076	876	-4.092	<0.00001
area2	-0.132	0.026	876	-5.122	<0.00001
area3	-0.064	0.014	876	-4.471	<0.00001
area4	-0.089	0.036	876	-2.449	0.01450
area5	-0.013	0.010	876	-1.280	0.20100
area6	-0.053	0.036	876	-1.472	0.14140
area7	0.012	0.030	876	0.406	0.68510
area8	-0.057	0.009	876	-6.420	<0.00001
area9	-0.021	0.007	876	-3.001	0.00280
area10	-0.013	0.008	876	-1.608	0.10810
area11	-0.010	0.006	876	-1.581	0.11420
area12	0.011	0.006	876	1.845	0.06530

area13	0.006	0.005	876	1.213	0.22540
area14	-0.014	0.017	876	-0.827	0.40830
area15	0.006	0.015	876	0.398	0.69040
area16	-0.008	0.008	876	-1.047	0.29560
area17	-0.002	0.003	876	-0.809	0.41900
area18	0.003	0.002	876	1.575	0.11570
area19	0.021	0.012	876	1.711	0.08740
area20	-0.003	0.004	876	-0.972	0.33120

Table 4: Estimated coefficients : **Zone G**.

Coefficient	Estimate	Std. error	df	t-value	$Pr(>\|t\|)$
Intercept	0.075	0.039	1241	1.930	0.05380
year1	0.047	0.062	1241	0.755	0.45040
year2	0.030	0.028	1241	1.087	0.27720
year3	-0.030	0.025	1241	-1.202	0.22960
year4	-0.065	0.044	1241	-1.476	0.14030
year5	0.040	0.040	1241	0.983	0.32590
year6	-0.009	0.011	1241	-0.880	0.37880
year7	-0.009	0.008	1241	-1.096	0.27320
year8	-0.015	0.006	1241	-2.459	0.01410
year9	-0.020	0.005	1241	-3.753	0.00020
year10	-0.032	0.005	1241	-6.324	<0.00001
year11	-0.014	0.007	1241	-1.930	0.05390
year12	-0.035	0.004	1241	-7.801	<0.00001
year13	-0.040	0.004	1241	-11.364	<0.00001
year14	-0.018	0.003	1241	-5.152	<0.00001
year15	-0.015	0.003	1241	-4.708	<0.00001
year16	-0.021	0.003	1241	-8.017	<0.00001
year17	-0.021	0.002	1241	-8.992	<0.00001
year18	-0.017	0.002	1241	-7.984	<0.00001
year19	-0.010	0.002	1241	-4.066	0.00010
area1	0.248	0.110	1241	2.246	0.02490
area2	-0.118	0.051	1241	-2.296	0.02180
area3	-0.088	0.023	1241	-3.829	0.00010
area4	-0.028	0.013	1241	-2.178	0.02960
area5	-0.020	0.010	1241	-2.038	0.04170
area6	-0.015	0.024	1241	-0.626	0.53110

area7	-0.021	0.013	1241	-1.629	0.10350
area8	-0.004	0.016	1241	-0.278	0.78140
area9	0.018	0.016	1241	1.169	0.24260
area10	-0.019	0.005	1241	-4.111	<0.00001
area11	-0.011	0.011	1241	-0.955	0.33990
area12	-0.003	0.003	1241	-0.871	0.38380
area13	-0.001	0.008	1241	-0.069	0.94500
area14	-0.030	0.018	1241	-1.630	0.10340
area15	0.009	0.011	1241	0.843	0.39910
area16	0.002	0.003	1241	0.631	0.52780
area17	-0.015	0.016	1241	-0.926	0.35480
area18	0.002	0.004	1241	0.511	0.60930