Western and
Central Pacific
Fisheries
Commission

SCIENTIFIC COMMITTEE
FOURTH REGULAR SESSION

11-22 August 2008
Port Moresby, Papua New Guinea

THE USE OF PRINCIPAL COMPONENTS ANALYSES TO ASSIST IN SELECTING VARIABLES TO INCLUDE IN A CATCH RATE STANDARDISATIONS

WCPFC-SC4-2008/ME-IP-1

Brett Molony [1], Kathy Sisior [2]

[1] Oceanic Fisheries Programme, Secretariat of the Pacific Community, Noumea, New Caledonia
[2] Oceanic Fisheries Management Section (OFMS), Bureau of Marine Resources (BMR), Republic of Palau

Abstract

Catch rate standardisations are critical to the development of indices of biomass for inclusion in stock assessment models. However, a potentially large number of operational and oceanographic variables are able to be considered to be included in a standardisation model. We use principal component analyses (PCAs) to demonstrate the utility of this approach to assist the analyst in identifying variables to include in catch rate standardisations.

Introduction

Constructing a standardised catch per unit effort (CPUE) series is critical to the development of indices of abundance in stock assessment models (Hoyle et al. 2007). In addition, standardisations of CPUE data are useful to understand the key variables that influence catch rates at smaller scales, potentially down to levels of EEZs or sub-EEZs of fishing fleets.

Standardisations of CPUE data have been undertaken using a range of modelling approaches (e.g. GLMs, GAMs) (Hoyle et al. 2007). These standardisations have generally been undertaken using a wide range of operational parameters (e.g. hooks per basket, set type, latitude, longitude, temporal variables (years, months, quarters)) that are available within a data set. For example, for some of the distant water fishery data that are supplied on 5° x 5° by month basis, latitude, longitude, year-month and hooks between floats (HBF) have been used to derive standardised catch rate indices to provide temporal series of relative biomass within stock assessment models (e.g. Langley et al. 2007). The implicit assumptions made by the supplying country and/or the analyst is that variables included in a standardisation represent the key variables that need to be included in a model.

A wide range of oceanographic variables have also been used to standardise catch rates (e.g. sea surface temperature, temperature at depth, current variables, salinity, altimetry, concentration of chlorophyll) (Hoyle et al. 2007). Other fishery variables (e.g. catches in the previous month of the target species, catches of other species) and the presence of particular equipment on fishing vessels (e.g. presence of bird radar) have also been considered (Shono and Ogura 2000).

There are potentially two issues associated with variable selection as described above. Firstly, the wide range of potential variables that could be included in a standardisation model makes the identification of important variables difficult. Secondly, some variables are likely to be highly correlated with each other. For example, temperature is likely to strongly affect the distribution and local abundances of tunas and other species; however, a wide range of temperature variables may be considered and incorporated in standardisations (e.g. SST, temperature at depth, depth of different isotherms, differences in depth of different isotherms, monthly ranges of temperatures etc) and many correlations among temperature variables are likely to exist. Thus several variables in a model may have little additional explanatory power compared to a single variable in standardisations (i.e. they are redundant). The problem is deciding which particular variables should be considered and included in a particular CPUE standardisation. This is often left to the discretion of the analyst, using relevant literature in regard to species preferences (e.g. depths, temperature preferences) on which to base variable selection.

One way of identifying important variables is to examine and compare the effects of different combinations of variables and fits of the subsequent model to the data. For example, standardisation might compare a suite of different GLMs, comparing the fit of a range of variables and models by AIC and the amount of deviance explained. However, combinations of variables included in each GLM still rely on the judgement of the analyst.

One alternative is to use multi-variate data exploration methods, such as principal components analyses (PCAs), to assist the analyst in identifying variables to include in a catch rate standardisation model. PCAs take multivariate datasets and ‘re-plot’ data in multivariate space, creating new axes (Principal component axes, PC axes). The original data (catch rates) are then re-plotted against these new axes. Correlations of each variable on each PC axis are determined to identify in determining the location of each PC axis. The outputs of PCAs include the weighting of each variable on each new PC axis to determine which variables are most influential in describing the variability in catch rate data. The total amount of variability in the data set accounted for by each PC axis is also estimated (McGarigal et al. 2000).

PCAs are not statistical tests of significance but assist identifying which variables in a multivariate data set are contributing most to explaining the variability in a data set. In addition, PCAs assist in identifying which variables may be considered redundant (providing little or no additional explanatory power), by allowing the examination of the weightings of individual variables on each axis. If several similar variables (e.g. depth of a range of temperature isotherms) have a similar weighting on a single PC axis, then the analyst may consider including only the most heavily weighted variable of a particular axis to include in a subsequent catch rate standardisation. Ultimately, PCAs can assist identifying which variables are most influential in explaining the variation in a dataset and therefore which variables could be considered for inclusion in a subsequent catch rate standardisation via a standard approach (e.g. GLM).

This paper provides a simple example of how a PCA may be used to assist in selecting variables to be included in a subsequent catch rate standardisation. The PCA approach is compared to a range of GLM standardisations where combinations of variables are selected by the analyses. We limit our analyses to consider only the influence of oceanographic variables on longline catch rates of bigeye from the vicinity of the Palau EEZ, 1998–2006.

Methods

To test the applicability of a PCA approach to variable selection for catch rate standardisations, a range of readily available oceanographic variables (www.NOAA.NCEP) and longline catch and effort data for bigeye tuna were merged into strata of 2° x 2° of latitude and longitude. Four strata representing areas of relatively high levels of recent bigeye catch and effort from the vicinity of the Palau EEZ were selected for further analyses. GLMs for a range of models were undertaken and the fits to each model compared (Table 1).

A PCA using a similar range of variables was subsequently undertaken for one of these strata. The weighting of each variable on each PC axis was examined, and the variable with the highest weighting (correlation) on each of the first four PC axes were selected. A GLM using these four variables (one from each of the first four PC axes) was then run and the fit compared to the original GLMs. All analyses were undertaken in R.

Results

Initial GLMs

A total of 13 GLMs were undertaken (Table 1) to compare the amount of variation in bigeye catch rates attributable to different combinations of oceanographic variables. Deviance explained by each GLM were compared to assess the fit of each model. Each model included a range of variables for each oceanographic characteristic (e.g. mean monthly value, highest monthly value, lowest monthly value, monthly range). In addition, a GLM model including

15 common oceanographic variables that were considered in a PCA (Table 2) was also undertaken.

Temperature and altimetry variables explained the highest proportion of variation in catch rates of any groups of variables (35% and 44% of deviance explained, respectively), with chlorophyll, altimetry and salinity accounting for a similar amount of deviance in the model but a more complex model.

The full model (all oceanographic variables) accounted for more than 58% of deviations in monthly bigeye catch rates from these four strata but identified 14 significant oceanographic variables (Table 1), a complex model, likely to be over-parameterised. Examination of the diagnostics (AIC) of this GLM model revealed that three variables (i.e. the Reduced model, which included monthly deviations from mean altimetry, the average monthly depth of the 27ºC isotherm, and the average monthly depth of the 18ºC isotherm) plus the month term accounted for approximately 53% of deviations in monthly bigeye catch rates from these four strata. Including catches in the previous month improved the amount of deviance explained by both the full model and reduced model.

Table 1. GLM fits for each of the bigeye models examined. Deviance explained indicates how much of the deviance in the CPUE data was explained by each model option. GLMs from the PCAs are provided in the lower portion of the table. The number of variables included in each model includes the 'month' term.

Model	**r^2**	**Deviance explained (adjusted r^2)**
GLMs		
Month only	0.143	11.9 %
Month + currents variables	0.122	9.7 %
Month + altimetry	0.456	44.3 %
Month + temperature variables	0.381	35.2 %
Month + salinity	0.071	5.5 %
Month + chlorophyll	0.088	7.1 %
Month + chlorophyll + altimetry + salinity	0.466	44.8 %
Full model	0.667	58.7 %
Reduced model	0.596	53.0 %
Previous catches only	0.172	16.8 %
Month-strata + previous catches	0.333	23.9 %
Full model + previous catches	0.708	62.9 %
Reduced model + previous catches	0.609	54.5 %
GLMs including only the variables included in the PCA		
PCA variables	0.623	54.7 %
PCA GLMs		
Oceanographic variables model	0.549	47.1 %
Oceanographic variables model + previous catches	0.585	51.0 %

PCA

Fifteen oceanographic variables were included in a PCA (Table 2). A GLM including these variables explained more than 54% of deviance in monthly catch rates of bigeye tuna from the area examined (Table 1).

A PCA was undertaken on these oceanographic variables, using monthly catch rates of bigeye as a grouping variable. Catch rate data were grouped into four levels; Very low (VL), catch rates less than 50% of the mean (12.6 kg.hhooks^{-1}); Low (L), catch rates greater than 50% of the mean but less than the mean; high (H), catch rates greater than the mean but less than 1.5 times the mean; Very high (VH), catch rates greater than 1.5 times the mean.

Variances were scaled to unit variance to allow differences in the variance among variables to be validly compared (R Development Core Team, 2007). The first four PC axes accounted for more than 77% of the variance in the original data set (Table 2). PC axis 1 was mainly influenced by variables describing the thermal profile of the water column in the strata being considered and accounted for more than 32% of the total variance in the data set (Table 2). The depth of the 22 °C isotherm produced the highest correlation on PC axis 1.

Table 2. Variable weightings on the first four principal component axes from a PCA examining the influence of the mean monthly value of 15 oceanographic variables on bigeye catch rates in the area examined of the western WCPO, 1998–2006. Grey filled cells highlight correlations of variables with PC axes of greater than 0.35. Values underlined identify the highest weightings on each PC axis, identifying those variables that were used in subsequent GLMs. The proportion of variance explained by each PC axis is also provided, as is the cumulative variance and the percent of total variance explained with the addition of each of the first four PC axes.

Variable	**PC axis 1**	**PC axis 2**	**PC axis 3**	**PC axis 4**
Sea surface temperature	0.222	0.360	-0.084	-0.252
Chlorophyll a concentration	-0.020	-0.438	0.216	0.318
Altimetry deviation	0.188	0.134	0.024	0.586
Surface salinity	0.210	-0.383	0.037	-0.344
Depth of the 20°C isotherm	0.415	0.104	0.163	0.149
Depth of the 27°C isotherm	0.363	0.194	-0.014	-0.171
Depth of the 22°C isotherm	0.428	0.133	0.101	0.053
Depth of the 18°C isotherm	0.395	0.069	0.177	0.192
Strength of easterly current	-0.253	0.331	0.368	0.029
Strength of westerly current	-0.251	0.315	0.370	0.113
Strength of northerly current	0.204	-0.321	0.239	-0.112
Strength of southerly current	0.177	-0.076	0.360	-0.303
Chlorophyll a range	-0.042	-0.340	0.243	0.253
Altimetry range	-0.109	0.089	0.578	-0.161
Salinity range	-0.117	-0.049	0.154	-0.273
Proportion of variance	0.323	0.189	0.161	0.100
Cumulative variance	0.323	0.512	0.673	0.773
Variance explained (%)	32.3 %	51.2 %	67.3 %	77.3 %

Sea surface temperature (SST) had a strong positive weighting along PC 2. However, the strongest weighting along PC axis 2 was with chlorophyll concentration, which revealed a strong negative correlation with the axis (Table 2). Monthly altimetry range was most strongly weighted on PC axis 3. Current variables were also strongly weighted along PC axis

3. Mean monthly altimetry deviations weighted most strongly along PC 4. A biplot of the relationships between the variables and bigeye catch rates on PC axes 1 and 2 highlighted the strong weightings with these four variables (Figure 1), especially variables describing temperature at depths.

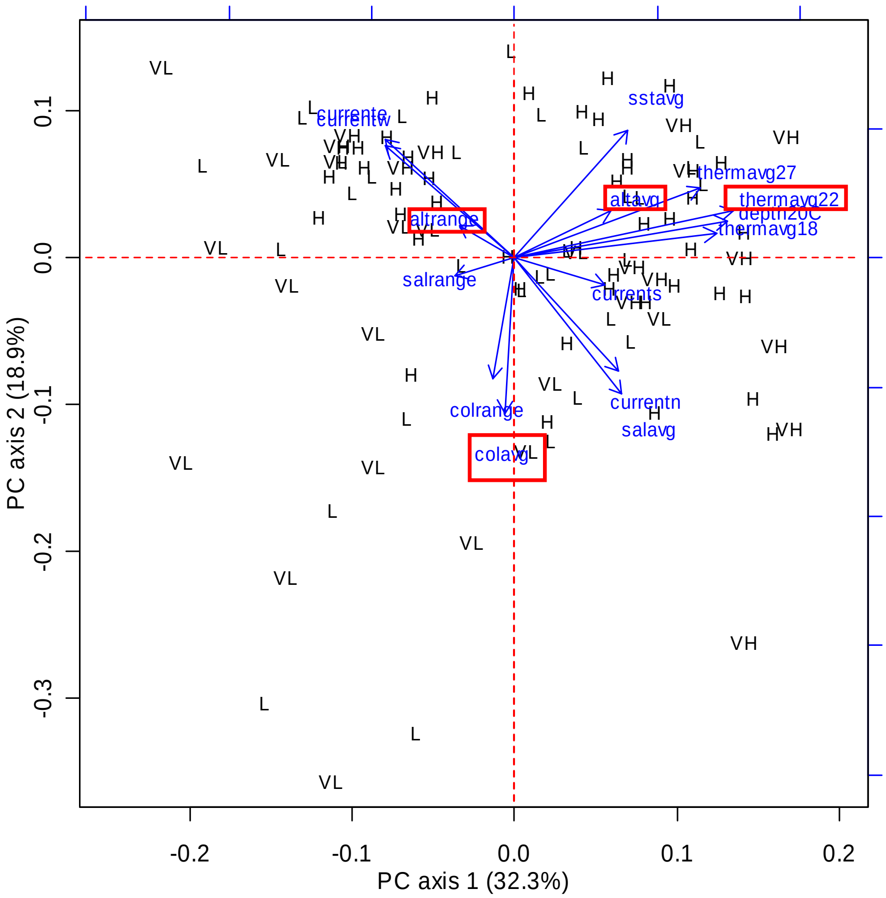

Figure 1. Biplot of relationships between oceanographic variables (blue arrows and labels) and bigeye catch rates (black labels) identified in a PCA. The biplot is of the first two PC axes, which account for more than 51% of variability in the multivariate data set. The length and vector of each arrow shows the relative weighting on the first two PC axes; longer arrows closer to each PC axis reveals stronger correlations. Catch rate labels (black letters): VL, catch rates less than 50% of the mean (12.6 kg.hhooks^{-1}); L, catch rates greater than 50% of the mean but less than the mean; H, catch rates greater than the mean but less than 1.5 times the mean; VH, catch rates greater than 1.5 time the mean. Variables in red boxes were used in subsequent GLMs.

The variables most heavily weighted on each of the first four PC axes (depth of the 22°C isotherm, chlorophyll a concentration, altimetry deviation from mean, monthly range of altimetry deviation) were included in a GLM to model monthly bigeye catch rates. The subsequent model accounted for more than 47% of the variability in monthly bigeye catch rates. Including catches in the previous month increased the fit to 51% of monthly variation in bigeye catch rates (Table 1).

Comparison of GLMs.

The four models examined in detail appeared to fit the long-term trend (1998–2006) in bigeye catch rates in each of the four strata well, with the exception that most models underestimated periods of relatively high bigeye catch rates in two of the four strata prior to 2000 (Figure 2).

The increasing nominal catch rates were well fitted by all four models, with the exception of periods of very high catch rates in one strata since 2004. The full (complex) model fitted the catch arte data best as expected, although the reduced model and PCA model also fitted the data from each strata well.

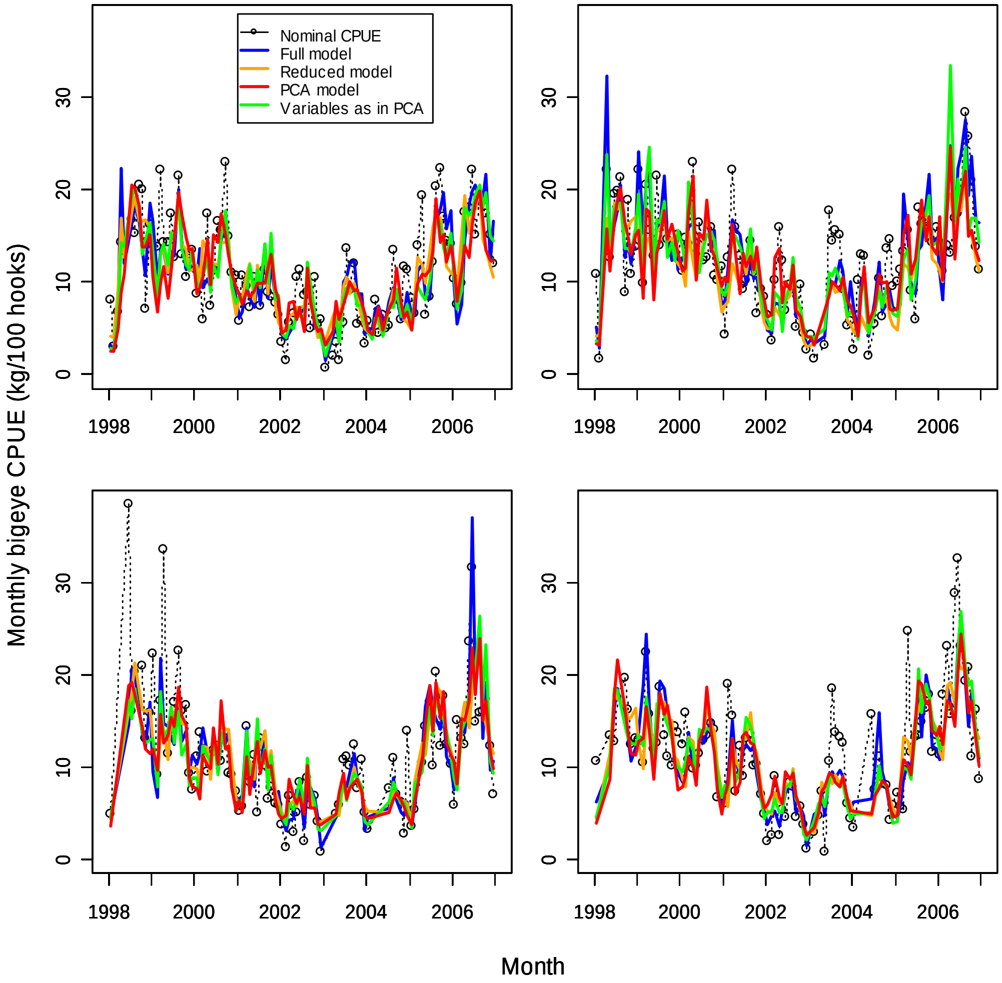

Figure 2. Comparison of fits of standardisations of bigeye catch rates (kg.hhooks^{-1}) from selected GLMs and GLMs using the outcomes of a PCA. Nominal catch rates are also plotted. Each figure represents an individual strata of 2° x 2°. "Variables as in PCA" represents the fit of a GLM which included all 15 variables included in the PCA (as listed in Table 2).

Discussion

All models examined in more detail appeared to fit the nominal bigeye catch rate data well (Figure 2). All models tended to reduce the influences of periods of relatively high or relatively low catch rates in all four strata. All models also included a downward trend in bigeye catch rates between 1998 and 2004, and subsequent increases in catch rates from 2005 onwards.

Catch rate standardisations from GLMs using the four variables selected from the PCA outputs explained a similar level of deviance to GLMs using a wider range of variables (Table 1). The 'Reduced model' GLMs produced a slightly better fit than the PCA GLMs, but involved the analyst selecting variables based on AIC outcomes available from GLM diagnostics. In contrast, the apriori decision to include only the variables with the highest weighting on each of the first four PC axes reduced the subjectivity of selecting only some variables identified as significant in a more complex GLM. This appears to be a more defensible approach to variable inclusion in a catch rate standardisation.

However, PCAs do have limitations. One of the major limitations is how many axes to interpret, and therefore, how many variables an analyst may consider including in a catch rate standardisation. For example, the current PCA generated 15 PC axes (Table 3). One approach to limit the number of PC axes to consider is to interpret the scree-plot of a PCA (Figure 3), one of the many diagnostics available (McGarigal et al. 2000, R Development Core Team, 2007). Two interpretations are possible. Firstly, the decision can be made on the point of inflection in the scree-plot. In the current analysis, the inflection occurs at PC axis 5 (Figure 3). In this case, it may be argued to analyse the data of the first five PC axes (i.e. include five variables in a subsequent standardisation); beyond this point, the amount of variance explained per additional PC axis declines (and a subsequent GLM becomes more complex for little additional benefit). Alternatively, only PC axes with unit variances greater than 1.0 could be interpreted (Figure 3), as has been done in this report (i.e. the first four PC axes).

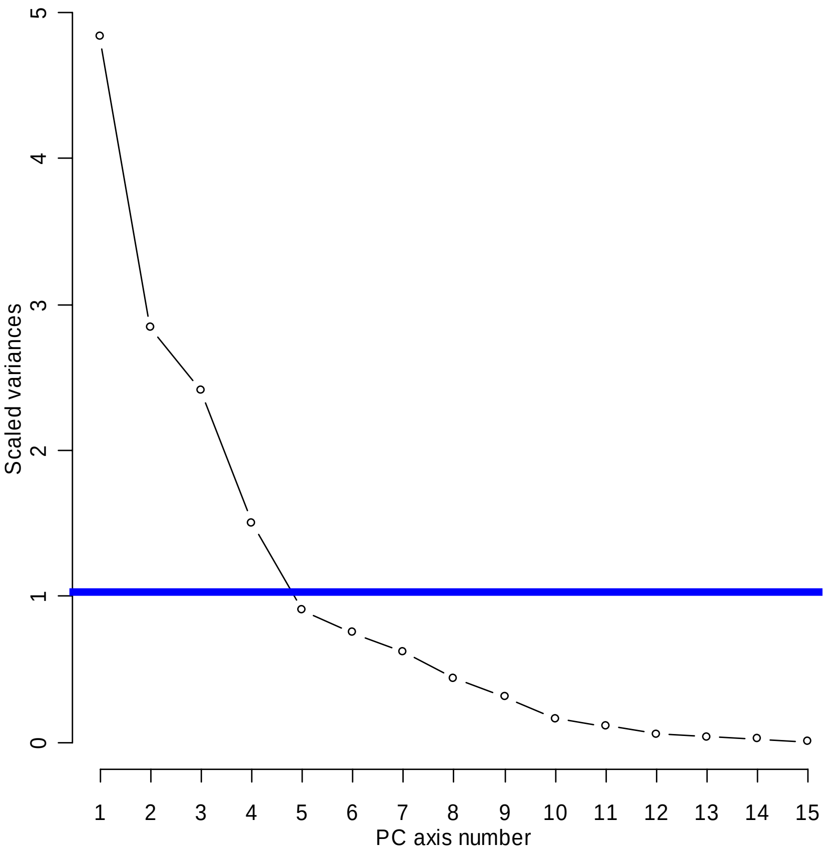

Figure 3. Scree-plot of the PCA described in the report. Variances have been scaled to unit variance. Points represent the amount of additional variance accounted for with the addition of each PC axis. The blue line indicates a scaled variance of 1.0.

The proportion of overall variance explained by the addition of each axis could also be used as a basis for how many axes should be considered (Table 3). For example, the first four axes all contribute 10% or more of total variance explained, with subsequent axes contributing less than 10%. In addition, an apriori decision on the total variance explained (cumulative proportion) could be used. Unfortunately, interpretation rules for PCAs are still poorly developed (McGarigal et al. 2000).

Table 3. The (unit) standard deviation (scaled variance) explained, proportion of variance and cumulative variance explained with the addition of each PC axis in the current PCA.

PC axis	Standard deviation	Proportion of variance	Cumulative proportion
PC 1	2.200	0.323	0.323
PC 2	1.685	0.189	0.512
PC 3	1.552	0.161	0.673
PC 4	1.226	0.100	0.773
PC 5	0.951	0.060	0.833
PC 6	0.866	0.020	0.883
PC 7	0.786	0.041	0.924
PC 8	0.659	0.029	0.953
PC 9	0.558	0.021	0.974
PC 10	0.399	0.011	0.985
PC 11	0.335	0.007	0.992
PC 12	0.242	0.004	0.996
PC 13	0.184	0.002	0.998
PC 14	0.147	0.001	1.000
PC 15	0.056	0.000	1.000

PCAs may be useful in a wide variety of data exploration situations, including but not limited to the identification of variables to be considered in catch rate standardisations using GLMs or other techniques. PCAs allow inclusion rules for variables to be established prior to analyses being undertaken. In addition, PCAs can incorporate as many variables as required and are highly robust to variables with a wide range of non-normal or varying distributions (McGarigal et al. 2000). Thus, the use of PCAs as a data exploration tool could be considered as a starting point in standardisations of catch rates involving a (potentially) large number of variables.

Nonetheless, both approaches (GLM only, or PCA and then GLM) identified that a similar set of variables (altimetry and temperature at depth) were important in influencing catch rates of bigeye from the four strata examined. The actual temperature-at-depth variable identified in the PCA (Table 2) varied from those in the reduced GLM model. From the PCA results , the weightings of the depth of temperature isotherms were similar (Table 2) and it is likely that the selection of an alternative temperature at depth variable to include in a GLM would have made little difference to the overall fit to the subsequent GLM.

Acknowledgements

We would like to thank Adam Langley, Don Bromhead and Simon Hoyle for comments on the draft document. We extend our thanks to the skippers and companies that provided logsheet data and the Statistics section of OFP for compiling the catch rate data.

References.

Hoyle, S., Bigelow, K., Langley, A., and Maunder, M. 2007. Proceedings of the pelagic longline catch rate standardization meeting. Information Paper ME-IP-1. Third Regular session of the Scientific Committee of the WCPFC, 13–24th August 2007, Honolulu, Hawaii, United States of America.

McGarigal, K., Cushman, S. and Stafford, S. 2000. Multivariate Statistics for Wildlife and Ecology Research. Springer Science + Business Media Inc, New York, United States of America.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. Reference Index. Version 2.6.0. R Foundation for Statistical Computing. http://www.gnu.org/copyleft/gpl.html.

Shono, H. and Ogura. M. 2000. The standardized skipjack CPUE, including the effect of searching devices, of the Japanese distant water pole and line fishery in the Western Central Pacific Ocean. ICCAT Skipjack Stock Assessment meeting, 1999. SCRS/99/59. ICCAT Collective Volume of Scientific Papers. **51(1):** 312–328.