Standardisation of CPUE data to account for variations in targeting in a mixed species linefishery: Operating model formulation

Henning Winker[a] , Sven E. Kerwath[b,a] and Colin G. Attwood[a]

[a]*Marine Research Institute, Zoology Department, University of Cape Town, Private Bag Rondebosch 7701, South Africa.*

[b]*Department of Agriculture Forestry and Fisheries, Private Bag X2, Roggebaai 8012, South Africa*

Introduction

Winker et al. (in press) developed a novel method to standardize multispecies catch-per-unit-effort (CPUE) data. This ‘Direct Principal Component’ method (DPC) uses continuous principle component scores (PCs), derived from a Principal Component Analysis of the catch composition data, as nonlinear predictor variables to adjust for the effect of temporal variations in targeting tactics that allocate effort towards particular species or species-complexes. The objective of this contribution is to test the performance of the DPC method by way of simulations.

a) Basic dynamics

An operating model was developed to generate multispecies CPUE records per fishing trip based on simulations of mixed-fisheries with two levels of complexity: (1) a simple mixed-fishery scenario (Fig. 1), comprising four target species that are unevenly distributed across two different habitats; and (2) a more complex mixed-fishery scenario (Fig. 2), comprising ten target species that are unevenly distributed across four different habitats. For illustrative purposes, we simulated mixed-fishery scenarios that broadly resemble the habitat associations of common target species in the South African linefishery.

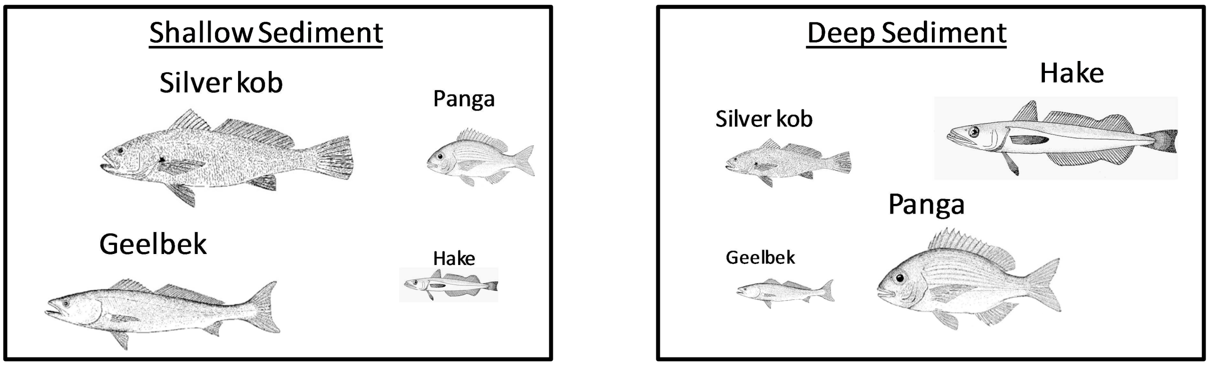

Fig.1 illustrates the mixed-fishery scenario for four target species that are unevenly distributed across two different habitat-types that are targeted by the fishery (H2.S4 scenario)

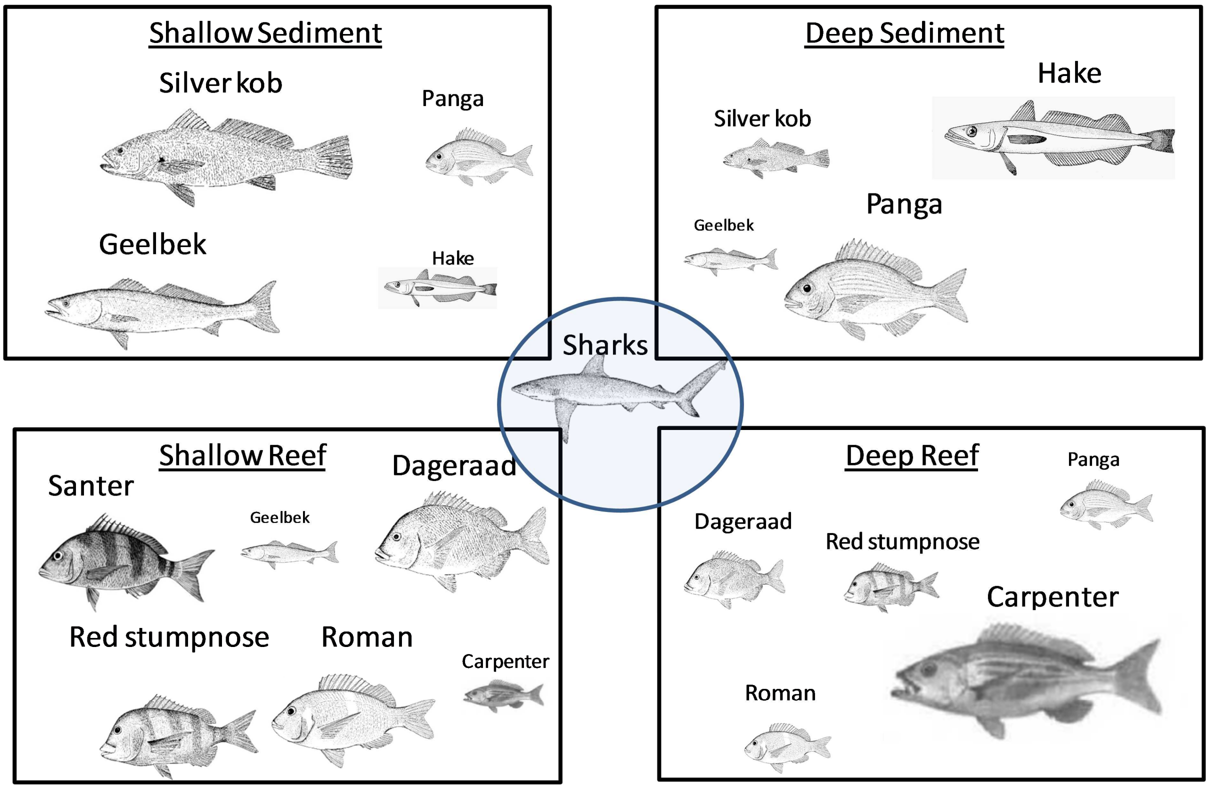

Fig.2 illustrates the complex mixed-fishery scenario for ten target species that are unevenly distributed across four different habitat-types that are targeted by the fishery (H4.S10 scenario)

The ‘true’ biomass for species *i* in year *y* was generated over a period of 20 years as a function of:

$$B_{i,y} = B_{i,1}e^{(r_i(y-1))} \qquad y = 1, 2, \ldots, 20. \qquad (1),$$

where $B_{i,1}$ is the biomass of species *i* at start of the time-series and r_i is the rate of increase (or decrease) for species *i*.

The use of CPUE as an index of abundance assumes that catch is proportional to the product of fishing effort and biomass:

$$C_{i,t} = q_i E_t B_i \qquad (2),$$

where $C_{i,t}$ is the catch of species *i* from trip *t, E_t* is the effort and q_i is the catchability for species *i* representing the fraction of biomass caught by expending one standard unit of effort. Here, we define one unit of effort as the effort expended during fishing trip *t*. Re-arranging this equation gives the relationship between CPUE and Biomass as:

$$\mathrm{CPUE}_{\mathrm{i,t}} = C_{i,t} / E_t = q_i B_i \qquad (3).$$

This relationship only holds if q_i is constant, which is almost certainly violated in mixed-fisheries that employ a variety of targeting tactics. Given that a particular targeting tactic will allocate effort towards a target species or species-complex, it will also influence the catchability of other species.

To simulate this effect, we assumed that the choice of targeting tactic is reflected by the choice of target habitat *j* during trip *t* and that each habitat is associated with a species-specific catchability $q_{\mathrm{i,j}}$ based on the conceptual consideration outlined in Stephens and MacCall (2004) and Winker et al. (in press). To promote a realistic generation of catch profiles for the individual trip level *t*, the CPUE from trip *t* for species *i* was assumed to be associated with a time-invariant capture probability $p_{i,j}$. The introduction of this parameter permits the presence/absence of species to vary across fishing trips, in the way that, for example, species that rarely frequent a particular habitat will also be less likely encountered in the catch that was taken from that habitat (Stephens and MacCall, 2004). The corresponding function used to generate $\mathrm{CPUE}_{t,i,y}$ is given by:

$$\mathrm{CPUE}_{t,i,y} = \begin{cases} 0 & \text{if } U(0,1) > p_{i,j} \\ q_{i,j} B_{i,y} & \text{otherwise} \end{cases} \qquad (4),$$

where U(0,1) denotes a random uniform number between 0 and 1. Note that if $p_{i,j} = 1$ for all species and habitats, equation (4) reduces to the familiar form:

$$\mathrm{CPUE}_{t,i,y} = q_{i,j} B_{i,y} \qquad (5).$$

The reason for not using eq. 5 to generate data is because it results in every species in a habitat being represented in every catch record for that habitat. This was deemed unrealistic as mono-specific catch records were commonly encountered in the database.

Model nomenclature

Simulation tests involving multi-species, multi-habitat operating models can quickly become mired in a vast number of permutations of scenarios and model formulations. In an attempt to simplify the study and facilitate comparisons we have devised the following system of nomenclature.

The successive terms in the following example string refer to (in order): Number of habitats (H*n*), '.' the number of species (S*n*), '.' the type of data transformation used in the PCA - two commonly used are 2nd and 4th root (abbreviated R2 and R4 respectively), '.' the number of PC axes used in the GAM (PC*n*). A PC0 value would imply no standardisation for targeting. Effort distribution, referring to one of the cases illustrated in effort models (E_n, where n refers to the habitat scenario specified below):

Example: the simplest model is H2.S4.R2.PC1.E_1

We do not aim to restrict the scope of model permutations, but rather propose a system to organise alternative runs. Workshop participants may propose alternative and additional simulations tests, but hopefully these could be accommodated in the existing framework

Four-species-two-habitats (H2.S4) scenarios

First, a simple mixed fishery is simulated, which targets four species in two different habitats (H2.S4). This scenario includes the species silver kob (KOB), geelbeek (GLBK), hake (HAKE) and panga (PANG) that are distributed across shallow- and deep water sediment habitats (Fig. 1). Silver kob and geelbek are abundant in shallow sediment habitats but are less common in deeper waters, whereas hake and panga represent the dominant target species in deepwater sediment habitats. The species-specific distributions across the two habitats are determined by 4 species (*i*) × 2 habitat (*j*) matrixes of $q_{i,j}$ and $p_{i,j}$ values, summarized in Table 1.

Table 1 summarizes choices of species- and habitat-specific $p_{i,j}$ and $p_{i,j}$ for the two-habitat-four-species scenarios (H2.S4). j = 1: shallow water sediment; j= 2: deep water sediment.

	Habitat *j*			
Species *i*	$p_{i,1}$	$p_{i,2}$	$q_{i,1}$	$q_{i,2}$
Silver kob	0.9	0.3	0.7	0.3
Geelbek	0.7	0.1	1.0	0.2
Hake	0.2	0.9	0.2	1.0
Panga	0.3	0.6	0.3	1.0

Two alternative effort scenarios were considered to simulate the distribution of fishing effort across habitats. This distribution was determined as the probability that habitat *j* is targeted in year *y*, $e_{j,y}$, such that $\sum_j e_{j,y} = 1$. The first effort scenario (E_1) simulates time-invariant probabilities $e_{j,y}$ and acts as a 'control' (Fig. 3a), while the second effort (E_2) scenario simulates a linear increase in $e_{j,y}$ for one habitat and a linear decrease for the other habitat (Fig. 3b).

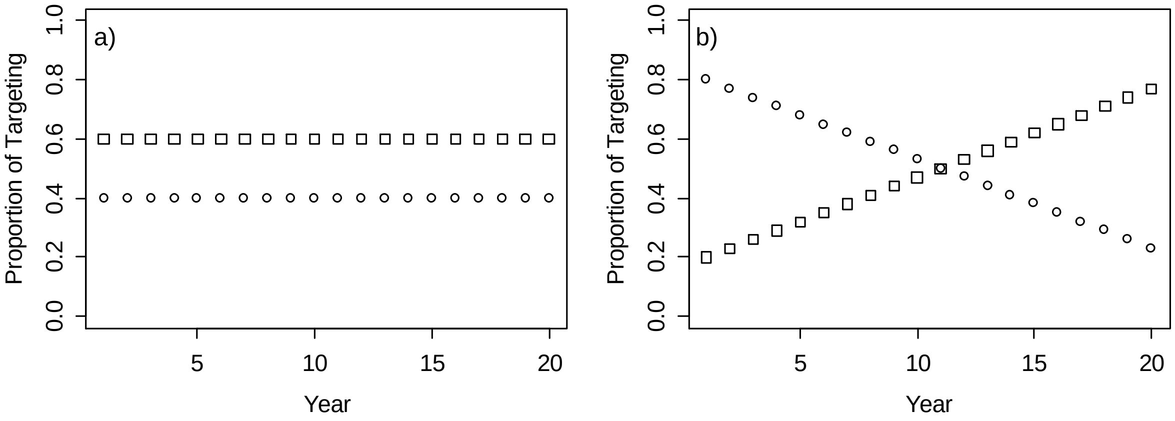

Fig. **3** Two effort allocation scenarios, E_1 and E_2 for H2.S4 models. Each symbol represents a particular habitat and the y-axes denote the probably that this habitat is targeted for any given year.

Ten-species-four-habitats (H4.S10) scenarios

In this scenario, the mixed-fishery was extended to ten target species, which are distributed across four different habitats (H4.S10): shallow- and deep water sediment and shallow- and deep water reef habitats (Fig .2). The species-assemblages of shallow- and deep water sediment habitats correspond to the first mixed-fishery scenario H4.S10 (Figs.1 and 2). The shallow water reef assemblage is dominated by roman (ROMN), dageraad (DRGD), red stumpnose (RSTM) and santer (SNTR), while carpenter (CRPN) represents the dominant target species over deep water reefs (Fig. 2). There is some distributional overlap among reef associated species. In addition, we introduced the group 'sharks' (SHRK), for which small catches are occasionally made in all four habitats. The species-specific distributions across habitats are determined by 10 species (*i*) × 4 habitat (*j*) matrixes of $q_{i,j}$ and $p_{i,j}$ values, summarized in Table 2.

Table 2 summarizes choices of species- and habitat-specific $p_{i,j}$ and $p_{i,j}$ for the four-habitat-ten-species scenarios (H4.S10). j=1: shallow water sediment; j=2: deep water sediment; j=3: Shallow water reef; and j=4: Deepwater Reef

	Habitat *j*							
Species *i*	$p_{i,1}$	$p_{i,2}$	$p_{i,3}$	$p_{i,4}$	$q_{i,1}$	$q_{i,2}$	$q_{i,3}$	$q_{i,4}$
Silver kob	0.9	0.3	0.0	0.0	0.7	0.3	0.0	0.0
Geelbek	0.7	0.1	0.0	0.1	1.0	0.2	0.0	0.1
Hake	0.2	0.9	0.0	0.0	0.2	1.0	0.0	0.0
Panga	0.3	0.6	0.0	0.0	0.3	1.0	0.0	0.0
Carpenter	0.0	0.0	0.8	0.2	0.0	0.0	1.0	0.2
Santer	0.0	0.0	0.1	0.8	0.0	0.0	0.1	0.7
Roman	0.0	0.0	0.2	0.7	0.0	0.0	0.2	0.8
Dageraad	0.0	0.0	0.2	0.4	0.0	0.0	0.1	0.6
Red stumpnose	0.0	0.0	0.3	0.5	0.0	0.0	0.1	0.5
Sharks	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.1

As in scenario H2.S4, we considered two alternative effort scenarios for the distribution of $e_{j,y}$, with the first (E_1) effort acting as a 'control' by simulating time-invariant probabilities $e_{j,y}$ (Fig 4a). The second effort scenario (E_2) simulates linear increases and decreases in $e_{j,y}$ for two habitats as well as a sudden switch $e_{j,y}$ for between the other two habitats (Fig. 4b).

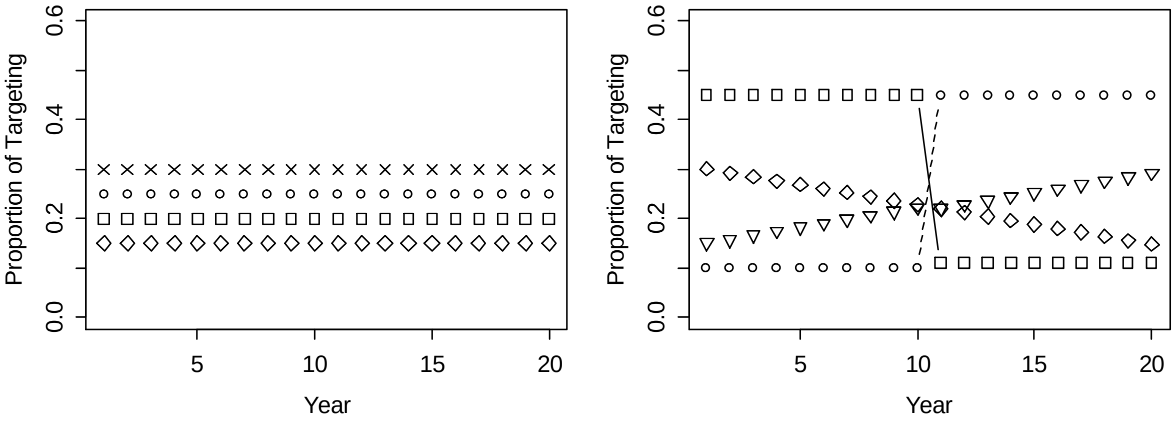

Fig. 4 Two effort allocation scenarios,E_1 and E_2, for H4.S10 models. Each symbol represents a particular habitat and the y-axes denote the probably that this habitat is targeted for any given year.

b) Data generation

A total of 100 simulation datasets were randomly generated for each of the four scenarios: (1) H2.S4.E_1, (2) H2.S4.E_2, (3) H4.S10.E_1 and (4) H4.S10.E_2. Each simulation dataset consisted of 500 trips per year and correspondingly a total 10000 trip records over the 20 years period.

The following randomization procedures were applied in order to generate the simulation datasets:

(1) Random biomass time series for each species i, $B_{i,y}$ (eq. 1), were generated by drawing random variants of r_i from uniform distribution with bounds at -0.1 and +0.1, U(-0.1, 0.1). Random biomass values at the start of the time series, $B_{i,1}$, were generated from a lognormal distribution as $B^*_{i,1} = 20000e^{(\varepsilon)}$ and $\varepsilon \sim N(0, 0.5^2)$ for the abundant species

silver kob, geelbek, hake, panga and carpenter; and as $B^{*}_{i,1} = 5000e^{(\varepsilon)}$ and $\varepsilon \sim N(0, 0.5^2)$ for the less abundant species santer, roman, dageraad, red stumpnose and sharks.

(2) The vectors $e_{j,y}$ that determine the probably for each habitat j being targeted in year y were randomly resampled without replacement, to vary the effort trends among habitats. Note that there were only two possible habitat × $e_{j,y}$ vector combination for the two-habitat scenarios but 24 possible habitat × $e_{j,y}$ vector combinations for the four-habitat scenarios.

(3) Random CPUE$_{t,i,y}$ deviates were drawn from a lognormal distribution associated with a CV ~ 20%, such that:

$$\text{CPUE}^{*}_{t,i,y} = \begin{cases} 0 & \text{if } U(0,1) > p_{i,j} \\ q_{i,j} B_{i,y}\, e^{(\varepsilon)} & \text{otherwise} \end{cases} \quad \text{and} \quad \varepsilon \sim N(0, 0.2^2) \qquad (6).$$

C) Standardization models

The simulated CPUE$_{t,i,y}$ data are standardised by applying the ‘Direct Principal Component method (DPC; Winker et al., in press). This method was developed on the premises that continuous principal component scores (PCs), derived from a Principal Component Analysis of the catch composition data, can be used as non-linear predictor variables for targeted effort within a Generalized Additive Model (GAM) framework (Winker et al., in press). The performance of this method is tested by comparing standardized CPUE indices with corresponding nominal CPUE indices.

The first step of the DPC method entails applying a PCA to a multidimensional CPUE$_{t,i,y}$ matrix. For this purpose, a data matrix only comprising CPUE$_{t,i,y}$ records was extracted from the simulation dataset. The CPUE$_{t,i,y}$ records were standardized into relative proportions by weight to eliminate the influence of catch volume and then square-root (R2) or fourth-root (R4) transformed to allow less dominant target species to contribute to the similarity among catch compositions and to shift the source of information away from raw abundance. Example results of PCAs are illustrated for scenario H2.S4.R4.E$_2$ (Fig. 5a) and scenario H4.S10.R4.E$_2$ (Fig .5b). In the final step of the dataset preparation, the first four PC-axes were directly aligned with the records in the original datasets for subsequent use as covariates in the GAM analysis.

Given that $p_{i,j}$ was assumed to be time-invariant for the purpose of this simulation study, we excluded records with zero CPUE$_{t,I,y}$ values for any species i under assessment before applying the species-specific CPUE standardization models. Note that this approach is equivalent to setting $p_{i,j}$ = 1 in delta-X (or hurdle) model formulation and does not introduce bias in the CPUE index in cases where $p_{i,j}$ is time-invariant.

Nominal CPUE indices were derived by applying a simple general linear model of the form:

$$\ln(CPUE_{t,i,y}) = \alpha + Y + \varepsilon \qquad (7),$$

where α is the intercept, Y denotes the categorical variable year and ε is the error term with $\varepsilon \sim N(0, \sigma^2)$. A bias-corrected estimate of expected mean CPUE for species i and year y is then given by:

$$\overline{\text{CPUE}}_{y,i} = \exp(\hat{\mu}_y + \frac{\hat{\sigma}^2}{2}) \qquad y = 1, 2,\ldots, 20 \qquad (8),$$

where $\hat{\mu}_y$ the expected mean ln(CPUE) for year y and $\hat{\sigma}^2$ is the estimated model standard deviation (residual standard error)

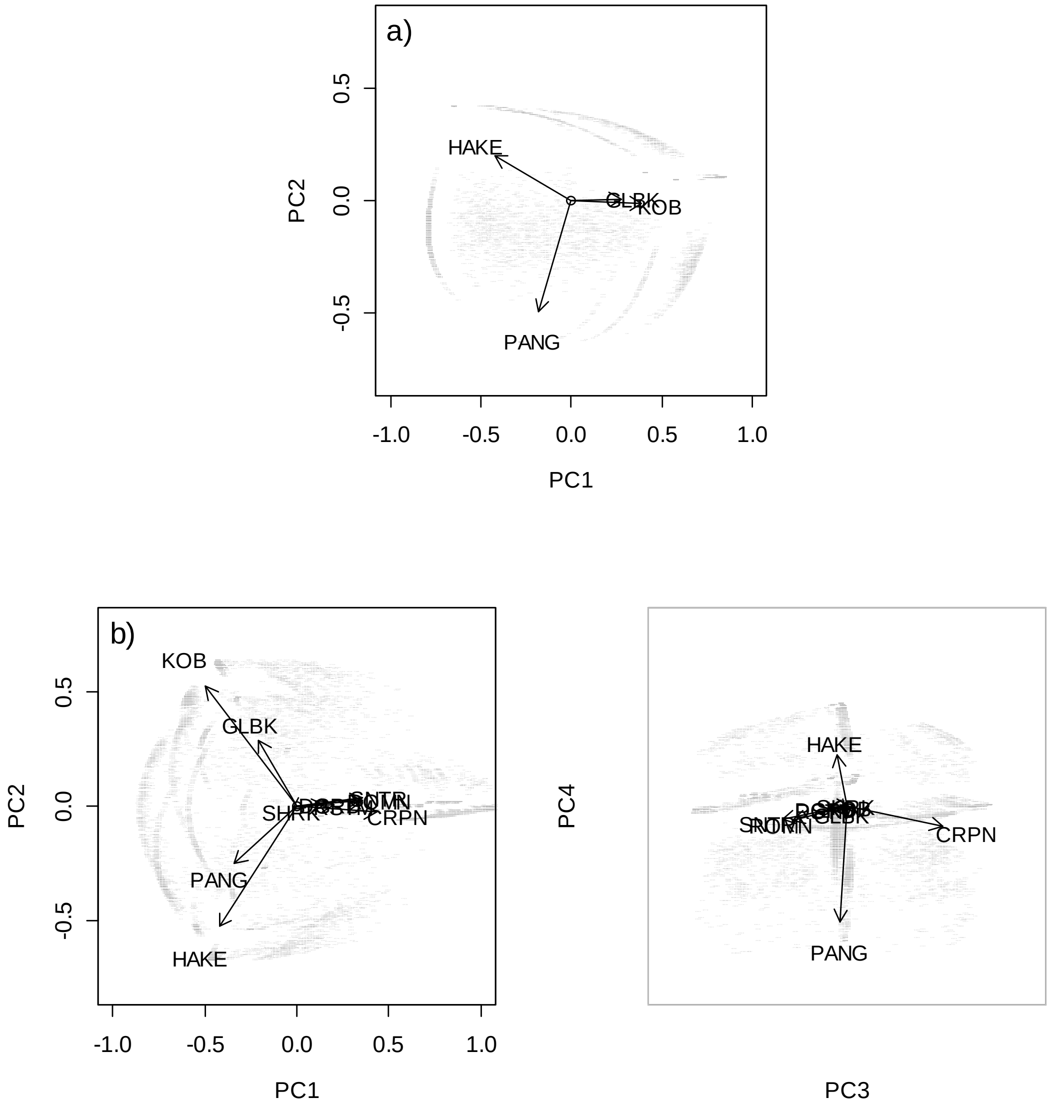

Fig. 5 . Biplots illustrate examples of simulated results from the principal components analysis (PCA) of fourth-root transformed catch composition matrixes for a) the first two principal component (PC) axes for scenario H2.S10.R4.E$_2$ and for (b) the first two PC-axes and c) PC-axes three and four for scenario H4.S.R4.E$_2$.

Conceptually, the number of Principal Component scores required to correctly separate the species assemblages that are associated with each target habitat *j* is given by the total number habitats minus one (see Fig 6). The GAMs for all two-habitat scenarios (H2.S4.R2.E$_1$, H2.S4.R2.E$_2$, H2.S4.R4.E$_1$ and H2.S4.R4.E$_2$) therefore included only the first principal component score (PC1) as non-linear predictor variable:

$$\ln(CPUE_{t,i,y}) = \alpha + Y + s(\mathrm{PC1}) + \varepsilon \qquad (9),$$

where *s*() denotes a thin plate regression spline smoother function, for which the maximum number of knots was limited to *k*≤5 in order to reduce the risk of 'over-fitting'.

The inclusion of first three PC scores (PC1-PC3) should theoretically produce the least biased CPUE indices for the four-habitat scenarios (H4.S10.R2.E$_1$, H4.S10.R2.E$_2$, H4.S10.R4.E$_1$ and H4.S10.R4.E$_2$). However, to examine the sensitivity of the results to under- or over-representation of non-linear PC predictors, we additionally formulated alternative GAMs that either only included PC1 and PC2 or that included the first four PC scores (PC1-PC4) , such that:

$$\ln(CPUE_{t,i,y}) = \alpha + Y + s(\mathrm{PC1}) + s(\mathrm{PC2}) + \varepsilon \qquad (10)$$

$$\ln(CPUE_{t,i,y}) = \alpha + Y + s(\mathrm{PC1}) + s(\mathrm{PC2}) + s(\mathrm{PC3}) + \varepsilon \qquad (11)$$

$$\ln(CPUE_{t,i,y}) = \alpha + Y + s(\mathrm{PC1}) + s(\mathrm{PC2}) + s(\mathrm{PC3}) + s(\mathrm{PC4}) + \varepsilon \qquad (12).$$

A bias-corrected estimate of expected standardized CPUE for species *i* and year *y* is given by:

$$\overline{\mathrm{CPUE}}_{i,y}(\overline{\mathbf{X}}^{\mathrm{T}}\hat{\beta}) = \exp(\hat{\mu}_y + \frac{\hat{\sigma}^2}{2}) \qquad y = 1, 2,\ldots, 20 \qquad (13),$$

where $\overline{\mathbf{X}}$ denotes means of the PC score vectors, $\hat{\beta}$ is the vectors of estimated coefficients and $\hat{\mu}_y$ the expected standardized ln(CPUE) for year *y*.

D) Performance measures

The performance of the DPC standardization models was evaluated in terms of the ability to accurately estimate r_i in comparison to the nominal CPUE indices. Estimates $\hat{r}_i$ were obtained from a simple linear regression of the form:

$$\ln(\overline{\text{CPUE}}_{i,y}) = \alpha + \hat{r}_i y \qquad y = 1, 2, \ldots, 20 \qquad (14).$$

The scenarios H2.S4.E$_1$ and H4.S10.E$_1$ act as control, for which $\hat{r}_i$ estimated from the nominal CPUE indices is expected to be unbiased. Measures of the Relative Error (RE) and the Absolute Relative Error (ARE) were used to summarize the estimation performance of $\hat{r}_i$ relative to the 'true' values r_i that governs $B_{i,y}$ (Ono et al., 2012). The Relative Error is a measure for the presence of systematic error and determines the overall tendency to overestimate or underestimate the true value of r_i :

$$RE_{i,h} = \frac{\hat{r}_{i,h} - r_{i,h}}{r_{i,h}} \qquad (15),$$

where $RE_{i,h}$ is the Relative Error from the h^{th} simulation and $\hat{r}_{i,h}$ is the estimate of r_i h^{th} simulation. A positive value for the MRE indicates that the standardization model tends to overestimate the true value of r_i and a negative value means the opposite.

The Absolute quantifies the average model precision and therefore provides a relative estimate for the goodness-of-the-fit:

$$ARE_{i,h} = \left| \frac{\hat{r}_{i,h} - r_i}{r_i} \right| \qquad (16).$$

Smaller values of the ARE$_{i,h}$ mean that $\hat{r}_{i,h}$ was estimated closely to r_i.

Relative Errors and Absolute Relative Errors from 100 simulation runs are presented in the form of boxplots for each scenario Hn.Sn.Rn.E$_n$ by species. Absolute Relative Errors are summarized over all species as Median Absolute Relative Error (MARE) for each scenario as:

$$\text{MARE} = \text{median}\left(\left| \frac{\hat{r}_{i,h} - r_{i,h}}{r_{i,h}} \right| \right) \qquad (17).$$

References

Ono, K., Punt, A.E., Rivot, E., 2012. Model performance analysis for Bayesian biomass dynamics models using bias, precision and reliability metrics. Fish. Res. 125, 173-183.

Stephens, A., MacCall, A., 2004. A multispecies approach to subsetting logbook data for purposes of estimating CPUE. Fish. Res. 70, 299-310.

Winker, H., Kerwath, S.E., Attwood, C.G., in press. Comparison of two approaches to standardize catch-per-unit-effort for targeting behaviour in a
multispecies hand-line fishery. Fish. Res.