



Canadian Journal of Fisheries and Aquatic Sciences  
Journal canadien des sciences halieutiques et aquatiques

# Target-based catch-per-unit-effort standardization in multispecies fisheries

Journal:	<i>Canadian Journal of Fisheries and Aquatic Sciences</i>
Manuscript ID	cjfas-2016-0460.R2
Manuscript Type:	Article
Date Submitted by the Author:	26-Apr-2017
Complete List of Authors:	Okamura, Hiroshi; National Research Institute of Fisheries Science, Fisheries Research Agency, Japan, Morita, Shoko; Hokkaido National Fisheries Research Institute Fisheries Research Institute, Higher Trophic Levels Group Funamoto, Tetsuichiro; Hokkaido National Fisheries Research Institute, Fisheries Research and Education Agency Ichinokawa, Momoko; National Research Institute of Far Seas Fisheries Eguchi, Shinto; Tokei Suri Kenkyujo Tachikawa Shisetsu
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	catchability, CPUE standardization, mixture regression model, targeting

SCHOLARONE™  
Manuscripts

Title: Target-based catch-per-unit-effort standardization in multispecies fisheries

Running title: Target-based CPUE standardization

Hiroshi Okamura (corresponding author)

National Research Institute of Fisheries Science, Fisheries Research and Education  
Agency, 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan.  
tel: +81-45-788-7514, fax: +81-45-788-5004, email: okamura@fra.affrc.go.jp

Shoko H. Morita

Hokkaido National Fisheries Research Institute, Fisheries Research and Education  
Agency, 2-2 Nakanoshima, Toyohira-ku, Sapporo, Hokkaido 062-0922, Japan.

Tetsuichiro Funamoto

Hokkaido National Fisheries Research Institute, Fisheries Research and Education  
Agency, 116 Katsurakoi, Kushiro, Hokkaido 085-0802, Japan.

Momoko Ichinokawa

National Research Institute of Fisheries Science, Fisheries Research and Education  
Agency, 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan.

Shinto Eguchi

The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562,  
Japan.

**Abstract:**

Standardized catch per unit effort (CPUE) is a fundamental component of fishery stock assessment. In multispecies fisheries, catchability can differ depending on which species is being targeted, and so the yearly trend extracted from the standardized CPUE is likely to be biased. We have, therefore, developed a method for predicting the unobserved variable related to targeted species from among multispecies composition data using a mixture regression model for the transformed residuals. In contrast to traditional methods, the proposed method predicts the target variable in CPUE standardization without removing a subset of the data. Keeping the entire dataset avoids information loss, and so CPUE standardization with the predicted target variable should yield an unbiased estimate of the yearly trend. Simple simulation tests demonstrate that our method outperforms traditional methods. We illustrate the use of our method by applying it to CPUE data on arabesque greenling (*Pleurogrammus azonus*) caught in multispecies trawl fisheries in Hokkaido, Japan.

**Keywords:** catchability, CPUE standardization, mixture regression model, targeting

## Introduction

Relative abundance (abundance index) is among the most fundamental pieces of data in fishery stock assessment and management (Hilborn and Walters 1992). The abundance index, which is generally assumed to be proportional to abundance (Maunder and Punt 2004), is used to tune the population dynamics model in many stock assessments and to monitor stock status for conservation and management (Hilborn and Walters 1992, Walters and Martell 2004, Geromont and Butterworth 2015, Tu et al. 2015).

An abundance index should, ideally, be fishery-independent. However, survey data to generate fishery-independent indices are often difficult to collect and involve substantial cost. Therefore, catch per unit effort (CPUE) calculated from commercial fisheries data is widely used for stock assessment (Maunder and Punt 2004), often in the absence of survey abundance indices. However, commercial fishers are selective in fishing, through non-random searches. Consequently, commercial CPUE is not typically proportional to abundance when a naïve method (such as raw catch rate) is used (Walters 2003). Because of this, a pure yearly trend that is representative of the abundance index must be extracted by adjusting (or removing) various factors (such as season and area) that change over the years and bias the raw CPUE trend. This operation is termed CPUE standardization (Maunder and Punt 2004), and it is typically achieved by the use of generalized linear models (GLMs) and their variants (GLMMs and GAMs) (Zuur et al. 2009).

Many fisheries catch multiple species simultaneously. When fishers target a particular species in a fishing trip, they may use a specific fishing operation that would not be optimal for catching other species. Therefore, the catchability coefficient should depend on target species. A variable indicating which species is targeted in a specific fishing

operation can be used as an explanatory variable in GLMs; however, target information is usually not available. We refer to this targeting behavior as “fishing tactics” in this paper.

The problem of adjusting the raw data can be tackled with directed CPUE, as described by Biseau (1998). The basic idea of directed CPUE is that the compositional data on catch or effort from multispecies fisheries can be used as surrogates for the unobserved variable (i.e., the fishing tactics). Stephens and MacCall (2004) use logistic regression for multispecies presence–absence data to predict the probability that the target species will be present.

One problem with these approaches is that they require the extraction of a subset of the data (“subsetting”). This process loses information and occasionally results in abundance indices with poor or insufficient precision. Moreover, such approaches do not allow comparisons among models before and after subsetting, because the datasets are incommensurate. In such cases, when we have two different yearly trends, how do we know which one should be used or believed?

Some authors have recently proposed alternative methods that do not require subsetting the data, including the clustering method (He et al. 1997), the directed principal component analysis method (Winker et al. 2013, 2014), the spatial imputation method (Carruthers et al. 2011), the spatial dynamic factor analysis (Thorson et al. 2016), and the decision-tree approach (Campbell 2016). Although these methods potentially solve the problems caused by data subsetting, they introduce other issues or problems. First, some of these methods assume that the fishing tactics are reflected in the selection of fishing grounds. Although this assumption is usually reasonable, there will be cases in which the fishing tactics are independent of spatial selection. Second, certain methods directly infer

(latent) fishing tactics from multispecies catch or CPUE composition. However, because multispecies catch or CPUE information (and the estimates derived therefrom) are not entirely independent of the response CPUE, model selection criteria such as the Akaike information criterion (AIC: Burnham and Anderson 2002) are likely to be invalid and the precision of the estimated year-effect tends to be overestimated (Winker et al. 2014).

We propose a new method for predicting the abundance index for the targeted species in a specific fishing operation using the CPUE composition of a multispecies fishery. We call our method the directed residual mixture (DRM) model. The DRM approach does not require subsetting data and thus is expected to yield more precise results than traditional methods, such as directed CPUE (Biseau 1998). The proposed method also reduces the dependence between the response variable and the estimated variable related to fishing tactics, which allows extensive model selection, including AIC usage.

We provide the DRM algorithm in the next section. The method is then tested with simulated datasets and its practical use is demonstrated by CPUE standardization applied to the stock of arabesque greenling (*Pleurogrammus azonus*) in the northern part of Hokkaido, Japan. Then, we discuss the advantages and disadvantages of the DRM approach in comparison with both traditional data-subsetting methods and recently developed multispecies CPUE approaches.

## Methods

### Directed residual mixture model

First, we explain the importance of taking residuals. Suppose that CPUE is proportional

to abundance  $N$ , so that  $\text{CPUE} = q \times N$ , where  $q$  is fishing efficiency. When we take the logarithm of CPUE, we have

$$(1) \quad \log(\text{CPUE}) = \log(q) + \log(N).$$

We assume that fishing efficiency differs depending on which species is targeted in a fishing operation. When there are  $K$  fishing tactics with different fishing efficiencies, we have

$$(2) \quad \log(\text{CPUE}_k) = \log(q_k) + \log(N),$$

where  $k = 1, \dots, K$  is the subscript indicating fishing tactics. When  $q_i \neq q_j$  ( $i \neq j$ ), any statistical analysis that assumes  $q$  is constant will sometimes provide incorrect results. The difference in fishing efficiency will be reflected in the multispecies CPUE composition, but direct comparisons of the raw CPUEs of different species do not always provide information on different fishing efficiencies. Here, we assume that differences in targeting lead to differences in fishing efficiency. Therefore, we attach importance to the relative difference in CPUEs, rather than to the absolute difference in CPUEs.

We denote the abundance of species  $s$  ( $s = 1, \dots, S$ ) in year  $y$  and the fishing efficiency of species  $s$  for fishing tactics  $k$  ( $k = 1, \dots, K$ ) by  $N_{s,y}$  and  $q_{s,k}$ , respectively. As an example, we consider  $K = S = 2$ . We assume that  $N_{1,1} = N_{2,1} = 1000$ ,  $N_{1,2} = 400$ ,  $N_{2,2} = 200$ ,  $q_{1,1} = 0.1$ ,  $q_{2,1} = 0.01$ ,  $q_{1,2} = 0.05$ , and  $q_{2,2} = 0.02$  (Table 1). If we determine that species 1 is the target when  $\log(\text{CPUE})$  for species 1 is higher than that for species 2, then we would incorrectly conclude from a comparison of the raw  $\log(\text{CPUE})$  data that species 1 is always targeted. However, by using the residuals of the  $\log(\text{CPUE})$  data,  $\log(\text{CPUE}_{s,k}) - 0.5 \sum_k \log(\text{CPUE}_{s,k})$ , we correctly determine when species 1 and 2 are targeted (Table 1).

Because a normal distribution of the logarithm of CPUE is often used to standardize CPUE (Maunder and Punt 2004), we use it to demonstrate our method. The number of

species in a multispecies fishery is  $S$ , and the number of fishing tactics,  $K$ , is assumed to be no more than the number of species (i.e.,  $1 \leq K \leq S$ ). We assume that the variable related to fishing tactics is unobserved and interacts with other observed variables. In the  $i$ th fishing operation ( $i = 1, \dots, n$ ) for species  $s$ ,  $X_{s,i}$  denotes a vector of variables that excludes the variable related to fishing tactics and its interactions and  $Z_{s,i}$  denotes a vector of variables that includes the variable related to fishing tactics and its interactions. We then have the CPUE standardization model for species  $s$ :

$$(3) \quad \log(\text{CPUE}_{s,i}) = X_{s,i}^T \alpha_s + Z_{s,i}^T \beta_s + \varepsilon_{s,i},$$

where the first element of  $X_{s,i}$  is 1 corresponding to the intercept,  $\alpha_s$  and  $\beta_s$  are the regression parameter vectors for  $X_{s,i}$  and  $Z_{s,i}$ , respectively, and  $\varepsilon_{s,i}$  are independent and identically distributed  $N(0, \sigma^2)$  random variables. The exponentiated estimate of yearly trend is the abundance index of interest (Maunder and Punt 2004, Stephens and MacCall 2004). However, we usually do not observe the variable related to fishing tactics. Thus, the model we can fit to the data is

$$(4) \quad \log(\text{CPUE}_{s,i}) = X_{s,i}^T \alpha_s + \nu_{s,i},$$

where  $\nu_{s,i} \sim N(0, \eta^2)$ . The residual  $\hat{\nu}_{s,i} = \log(\text{CPUE}_{s,i}) - X_{s,i}^T \hat{\alpha}_s \approx Z_{s,i}^T \hat{\beta}_s + \hat{\varepsilon}_{s,i}$ , where  $\hat{\alpha}_s$  is the maximum likelihood estimate (MLE) for  $\alpha_s$ , should contain information on the variable related to fishing tactics. If  $\hat{\nu}_{s,i}$  is large and positive when species  $s$  is targeted (a large  $\hat{\nu}_{s,i}$  indicates a high level of fishing efficiency for species  $s$  in fishing operation  $i$ ), then the exponentiated residual divided by the sum of exponentiated residuals, which is the ratio of relative CPUE for species  $s$  to the total relative CPUE in the  $i$ th operation, written as

$$(5) \quad \hat{p}_{s,i} = \frac{\exp(\hat{\nu}_{s,i})}{\sum_{u=1}^S \exp(\hat{\nu}_{u,i})},$$

should also be large. Because the proportion data of the exponentiated residuals are



constituted from multiple fishing tactics with different fishing efficiencies and will show a specific pattern that depends on the variable related to fishing tactics, we assume that the logit transformation of  $\hat{p}_{s,i}$  ( $\hat{z}_{s,i} = \text{logit}(\hat{p}_{s,i}) = \log[\hat{p}_{s,i}/(1 - \hat{p}_{s,i})]$ ) has a normal (or Gaussian) mixture model of linear regressions with  $K$  components:

$$(6) \quad f(\hat{z}_{s,i}) = \sum_k \pi_k \phi(\hat{z}_{s,i}; \mu_{k,i}, \sigma_k^2),$$

where

$$\phi(\hat{z}_{s,i}; \mu_{k,i}, \sigma_k^2) = N(\hat{z}_{s,i}; \mu_{k,i}, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[-\frac{(\hat{z}_{s,i} - \mu_{k,i})^2}{2\sigma_k^2}\right]$$

is the probability density function of linear regression and  $\{\pi_k\}$  are the mixing proportions with the constraint  $\sum_k \pi_k = 1$ . The mixture model comprises linear combinations of distinct probabilistic distributions (Bishop 2006, McLachlan and Krishnan 2008). The parameter  $\mu_{k,i} = M_{k,i}^T \omega_k$  is the expectation given the fishing tactics  $k$ ,  $M_{k,i}$  is a vector of explanatory variables in which the first element is always 1 and the subsequent elements are related to observed explanatory variables,  $\omega_k$  is the regression coefficient, and  $\sigma_k$  is the standard deviation of the normal distribution for fishing tactics  $k$ . The intercept term is the main effect of fishing tactics  $k$ , and the coefficients on explanatory variables are the interactions between those variables and fishing tactics  $k$ . Because  $\hat{z}_{s,i}$  can be a mixture of nonlinear functions of linear predictors, nonlinear modeling such as GAM may be expected to perform better. However, we use a mixture of linear models as an approximation to ease calculation. The logit function is not the only option for transforming binary data into real numbers. In sensitivity tests, we also use the probit function and the complementary log-log function instead of the logit function.

The parameters are estimated by the expectation–maximization (EM) algorithm (Dempster et al. 1977, Bishop 2006, McLachlan and Krishnan 2008), which alternates

between two types of update steps (called E and M) to find maximum likelihood estimates in latent variable models, with the likelihood function  $\prod_{i=1}^n \Pr(\hat{z}_{s,i})$ . As a result, we obtain a variable that indicates whether the datum was targeted or bycaught, on the basis of the posterior probability of membership of the components of the mixture, given by

$$(7) \quad \hat{t}_{s,i} = \operatorname{argmax}_h \frac{\hat{\pi}_h f(\hat{z}_{s,i}; \hat{\mu}_{h,i}, \hat{\sigma}_h^2)}{\sum_{l=1}^K \hat{\pi}_l f(\hat{z}_{s,i}; \hat{\mu}_{l,i}, \hat{\sigma}_l^2)}.$$

The variable  $\hat{t}_{s,i} \in \{1, \dots, K\}$  is a categorical variable that indicates a fishing operation with a different catchability coefficient or different fishing tactics. Hereafter, we refer to this variable as the “target variable.”

To select the appropriate number of components  $K$  ( $1 \leq K \leq S$ ), we use the consistent AIC (CAIC; Bozdogan 1987, Claeskens and Hjort 2008):  $\text{CAIC} = -2 \log\text{-likelihood} + p(\log(n)+1)$ , where  $p$  is the number of parameters and  $n$  is the sample size. CAIC generally detects the true number of mixture components very well (Lukočienė and Vermunt 2010). The model with the lowest CAIC is selected as the best model when examining the number of components, testing from 1 to  $S$ . If the number of components selected is equal to one, then we infer that there is no targeting effect and  $\log(\text{CPUE})$  is unrelated to the target variable.

With the target variable in hand, we can fit a linear model to the  $\log(\text{CPUE})$  data of species  $s$ :

$$(8) \quad \log(\text{CPUE}_{s,i}) \sim N(\eta_{s,i}, \tau_s^2).$$

Here,  $\eta_{s,i} = X_{s,i}^T \alpha_s + \hat{Z}_{s,i}^T \beta_s$ ,  $\alpha_s$  and  $\beta_s$  are the regression coefficients,  $X_{s,i}$  is a vector of the intercept (i.e., 1) and explanatory variables other than the target variable,  $\hat{Z}_{s,i}$  is a vector of the target variable  $\hat{t}_{s,i}$  and its interactions with other variables ( $\hat{Z}_{s,i}$  does not include the intercept term), and  $\tau_s^2$  is the variance of the distribution of  $\log(\text{CPUE}_{s,i})$ .

Extracting the appropriate year effect from eq. 8 gives the standardized abundance

index, which takes the variable target effect into account. The composition of fishing tactics based on eq. 5 and  $f(\hat{z}_{s,i})$  in eq. 6 are scale-independent. An additional reason that we first take the residuals of the linear model for each species is that this approach eliminates the effects of variables other than the missing target variable and its interactions. This should lead to more informative patterns that discriminate among different fisheries with different catchabilities. In addition, taking residuals should reduce the influence of the response variable. Although we restrict the equations to normal models for both the mixture analysis and the linear regression, they can easily be extended to GLMs or their variants (Zuur et al. 2009).

### Simulation test for the new method

We examine the performance of the DRM approach using a simulation test with a simple linear model. The objective is to extract an unbiased yearly trend when the target variable is missing. For simplicity, we suppose that the fishery targets two species, 1 and 2. When the fishery targets species 1, the catchability coefficient for species 1 is high and that for species 2 is low, and vice versa. The indicator variable of target species is denoted  $z$ . This variable has a Bernoulli distribution (Pawitan 2001) with expectation  $p$ , which varies over the years to simulate the yearly transitions of targeted species:

$$z_i \sim \text{Bernoulli}(p_i),$$

where  $p_i = 1/\{1 + \exp[-(a + b \text{year}_i)]\}$ ,  $a$  and  $b$  are the intercept and slope of regression, respectively, and  $\text{year}_i$  is the variable representing the year of the fishing operation  $i$  ( $i = 1, \dots, n$ ). We introduce one additional variable in the simulation,  $e$ , which is related to differences in fishing areas, (e.g., a fishery can have different catchabilities

in eastern and western areas). The variable  $e$  has a formulation similar to that of the variable  $z$ :

$$e_i \sim \text{Bernoulli}(q_i),$$

where  $q_i = 1/\{1 + \exp[-(c + d\text{year}_i)]\}$ , and  $c$  and  $d$  are the intercept and slope of regression, respectively. The difference between the variables  $z$  and  $e$  is that the former is an unobserved or latent variable whereas the latter is an observed variable.

The linear trend models for two species, in which the intercepts of lines change depending on the unobserved variable  $z$  and the observed variable  $e$ , are then given by

$$y_{s,i} \sim N(\mu_{s,i}, \sigma_s^2),$$

where  $s$  denotes the species identifier ( $s \in \{1, 2\}$ ),  $y_{s,i} = \log(\text{CPUE}_{s,i})$ ,  $\mu_{s,i} = \alpha_{s,z_i,e_i} + \beta_s \text{year}_i$ , and  $\alpha_{s,z_i,e_i}$  varies depending on species  $s$ , the unobserved variable  $z_i$ , and the observed variable  $e_i$ . We assume  $\alpha_{s,z_i,e_i} = \delta_s \gamma_{(2-s)z_i+(s-1)(1-z_i)+1} (1 - \varepsilon_s \times e_i)$ , in which  $\delta_s$  is a scale parameter that determines the size of catches of species  $s$ ,  $\gamma_{(2-s)z_i+(s-1)(1-z_i)+1}$  is a catchability coefficient that depends on  $s$  and  $z_i$ , and  $\varepsilon_s$  is a parameter related to effort allocation in fishing area  $e$  for species  $s$ . For example, when  $\delta_1 = 1$ ,  $\delta_2 = 2$ ,  $\gamma_1 = 10$ ,  $\gamma_2 = 2$ ,  $\varepsilon_1 = 0.1$ , and  $\varepsilon_2 = 0.9$ , we have  $\alpha_{1,0,0} = 10$ ,  $\alpha_{1,1,0} = 2$ ,  $\alpha_{1,0,1} = 9$ ,  $\alpha_{1,1,1} = 1.8$ ,  $\alpha_{2,0,0} = 4$ ,  $\alpha_{2,1,0} = 20$ ,  $\alpha_{2,0,1} = 0.4$ , and  $\alpha_{2,1,1} = 2.8$ . Our objective is to estimate the yearly trend  $\beta_s$  without using direct information about the variable  $z$  related to the target species.

We prepared four scenarios with different combinations of parameters (Scenarios 1 to 4 in Table 2). We set the number of years to 10 and the annual sample size (i.e., the number of fishing operations) to 100. The total sample size for each species is thus  $n = 10 \times 100 = 1,000$ . The simulation was repeated 1,000 times under each combination of parameters (Table 2). An additional scenario in which the response variable is unrelated

to  $z$  (i.e.,  $\gamma_1 = \gamma_2$ ) was also examined (Scenario 0).

We did not consider interaction between the year and variable  $z$  in any of the scenarios because we cannot know in advance whether the yearly trend for  $z = 0$  or  $1$  is correct as the abundance index. When an interaction between the year and variable  $z$  exists, we must choose year trend $|z = 0$  or year trend $|z = 1$ . We will refer back to this issue in the discussion section.

The yearly trends for species 1 and 2 are estimated using the following models: 1) the complete data model (Z:  $z$  is known); 2) the incomplete data model (nZ:  $z$  is simply ignored); 3) Biseau's directed CPUE model (DC: Biseau 1998); 4) directed PCA (DPC: Winker et al. 2013, 2014); 5) the cluster analysis approach (CL: He et al. 1997); and 6) our proposed model (DRM:  $z$  is replaced by  $\hat{t}$  in eq. 7).

We used the model

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}z_i + \beta_{s,3}e_i + \beta_{s,4}(z_i \times e_i) + \varepsilon_{s,i}$$

for the complete data model, where  $year_i$  is a continuous variable,  $z_i \times e_i$  is the interaction between the target variable  $z_i$  and the area variable  $e_i$ ,  $\beta_{s,\cdot}$  is the regression coefficient, and  $\varepsilon_{s,i} \sim N(0, \sigma_s^2)$ . Because  $z_i$  is missing for the incomplete data model, we ignored  $z_i$  while fitting the incomplete data model:

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + \varepsilon_{s,i}.$$

In Biseau's directed CPUE (DC), the composition of multispecies CPUE,  $q_{s,t,i} = \exp(y_{s,t,i}) / [\exp(y_{1,t,i}) + \exp(y_{2,t,i})]$  ( $s = 1, 2; t = 1, \dots, 10$ ), was calculated first. Then, the CPUE was permuted according to the ordered species composition,  $q_{s,t,(i)}$ , for each year ( $q_{s,t,(1)} = \min\{q_{s,t,1}, \dots, q_{s,t,100}\} \leq q_{s,t,(2)} \leq \dots \leq q_{s,t,(100)} = \max\{q_{s,t,1}, \dots, q_{s,t,100}\}$ ). Next, the cumulative CPUE,  $CCPUE_{s,t,(i)} = \sum_{j=1}^i CPUE_{s,t,(j)}$ , was calculated by summing from the minimum datum to the  $i$ th datum for each year. We finally fitted

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + \varepsilon_{s,i}$$

for the data subsetted by the condition  $100 \times CCPUE_{s,t,(i)} / \sum_{j=1}^{100} CCPUE_{s,t,(j)} > X\%$  ( $X = 25, 50$ , and  $75\%$ ).

For DPC, we first applied principal component analysis to the fourth root of the relative multispecies CPUE composition data,

$$\left( \sqrt[4]{\frac{\exp(y_{1,i})}{\exp(y_{1,i}) + \exp(y_{2,i})}}, \sqrt[4]{\frac{\exp(y_{2,i})}{\exp(y_{1,i}) + \exp(y_{2,i})}} \right).$$

We then fitted the model using the predicted principal component scores as a covariate,

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + s(PC_i) + \varepsilon_{s,i},$$

where  $PC_i$  is the first principal component score and  $s(\cdot)$  denotes a thin plate regression spline smoother function (Winker et al. 2013, 2014). Because we have only two species in this simulation, we used only the first principal component score. The fourth root transformation of CPUE composition data and the number of principal component scores used here are those recommended in Winker et al. (2014). Although we tried to use  $s(e_i, PC_i)$  instead of the above specification, the subsequent results were almost the same, so we referred only to the results from the above model.

For DC, we applied hierarchical cluster analysis (Ward method) to the arcsine-square-root of the relative multispecies CPUE composition data (He et al. 1997). Although He et al. (1997) first applied non-hierarchical cluster analysis (K-means method) to reduce the number of independent data from 46961 sets to 2500 non-hierarchical clusters, we did not use non-hierarchical clustering here because our simulated data are relatively simple and the sample size (1000 in our simulation) is not large. We then fitted

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + \beta_{s,3}\hat{c}_{s,i} + \beta_{s,4}(e_i \times \hat{c}_{s,i}) + \varepsilon_{s,i}.$$

Here,  $\hat{c}_{s,i}$  is the number of hierarchical clusters which was fixed at two.

Finally, for DRM, we first calculated the residuals (eq. 4) from the linear model,  $y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + \varepsilon_{s,i}$ . We then calculated the species composition (eq. 5) from the exponentiated residuals, fitted the mixture regression model (eq. 6) for the transformed values of compositional data (applying the probit and complementary log-log transformations as alternatives to the logit transformation), and obtained the target variable  $\hat{t}_{s,i}$  from eq. 7. Finally, we fitted the model

$$y_{s,i} = \beta_{s,0} + \beta_{s,1}year_i + \beta_{s,2}e_i + \beta_{s,3}\hat{t}_{s,i} + \beta_{s,4}(e_i \times \hat{t}_{s,i}) + \varepsilon_{s,i}.$$

The yearly trends of species 1 and 2 from each model are compared with the true values  $(\beta_1, \beta_2) = (-1.5, -0.5)$  (Table 2). The simulation was conducted using the free statistical software R (R Core Team 2016). The source code of the program is available in the online Supplement A.

We also carried out a simulation in which the number of species was increased from two to four. The details of that simulation are provided in the online Supplement B.

### Application of the new method to arabesque greenling CPUE data

Arabesque greenling (*Pleurogrammus azonus*) is a local fish species caught mainly in Hokkaido, Japan. Roasted arabesque greenling is a popular food in Japanese-style taverns. Recently, the catch and body size of this species have declined, and some people are concerned that this trend is a symptom of overfishing (Morita 2015). Arabesque greenling prefer cold waters and spawn in autumn, primarily in the adjoining water of the Sea of Japan (Tu et al. 2015, Morita et al. 2015). Recent recruitment of arabesque greenling has been very low, probably due to the recent increase in sea temperature. Thus, environmental changes and high fishing pressure might have both substantially impacted

the species. In light of this possibility, careful monitoring of the stock status is necessary.

For CPUE standardization, we used catch and effort data for the northern Hokkaido stock of arabesque greenling caught by bottom trawling. Bottom-trawl CPUE data have been used previously for arabesque greenling stock assessments (Morita 2015). Three species (arabesque greenling, walleye pollock [*Gadus chalcogrammus*], and Pacific cod [*Gadus macrocephalus*]) are caught in the same fishery. Walleye pollock and Pacific cod, which are also important seafood sources in Japan, have exhibited changes in abundance that differ from those of arabesque greenling. Because the three species inhabit different (optimal) horizontal and vertical habitats, and the fishing tactics for each species are likely to be different (Sakurai et al. 2007, Hokkaido Fisheries Experiment Station 2010, Narimatsu et al. 2015), we suspected that the catchability of arabesque greenling has changed over time.

The catch and effort data for bottom trawling between 1997 and 2015 comprise 66,094 records, which we used for the analysis of the catch composition and fishing effort in conjunction with other explanatory variables (area and month). Because the CPUE dataset includes zeros, we used the delta lognormal model for CPUE standardization (Lo et al. 1992). We first fitted a binomial model

$$Z_{s,y,a,m} \sim \text{Bin}(1, p_{s,y,a,m}),$$

where  $Z_{s,y,a,m} = 1$  for  $\text{CPUE}_{s,y,a,m} > 0$  and  $Z_{s,y,a,m} = 0$  for  $\text{CPUE}_{s,y,a,m} = 0$ . The indices  $s$ ,  $y$ ,  $a$ , and  $m$  denote species, year, area (small strata set up for management, see Supplementary Fig. C1), and month, respectively, and  $\text{logit}(p_{s,y,a,m}) = \theta_s + \theta_{s,y} + \theta_{s,a} + \theta_{s,m}$ , where  $\theta$  are effects related to the subscripts. Because zeros are not frequent in our dataset, we used only the main effects (i.e., year, area, and month) for the binomial model to avoid a convergence problem. We then fitted the normal model for the log-



transformed positive CPUE data,

$$\log(\text{CPUE}_{s,y,a,m}) \sim N(\mu_{s,y,a,m}, \sigma_s^2),$$

where  $\mu_{s,y,a,m} = \mu_s + \mu_{s,y} + \mu_{s,a} + \mu_{s,m} + \mu_{s,y,a} + \mu_{s,y,m}$ . In this,  $\mu_{s,y,a}$  is the interaction term between year and area,  $\mu_{s,y,m}$  is the interaction term between year and month, and  $\sigma_s$  is the standard deviation. The model was fitted separately to the CPUE of each species (arabesque greenling, walleye pollock, and Pacific cod). Here, all explanatory variables are categorical.

We then calculated the deviance residual (Pawitan 2001) for each species,  $D(\mathbf{y}; \hat{\boldsymbol{\theta}}) = -2[L(\hat{\boldsymbol{\theta}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]$ , where  $L$  is the maximized log likelihood of the delta lognormal model,  $\hat{\boldsymbol{\theta}}$  is an MLE, and  $L(\mathbf{y}; \mathbf{y})$  is the saturated model. We then obtained the probability of arabesque greenling being targeted, using the exponentiated deviance residuals (eq. 5). The use of deviance residuals enables us to address zero data in a natural way. We fitted the mixture linear regression model to the logit transformation of the probability (eq. 6). The number of components in the mixture model was selected according to the CAIC criterion. We predicted the target variable ( $\hat{t}$ ) from the posterior probability of membership of the components of the mixture obtained by the EM algorithm (eq. 7). Finally, we fitted the binomial model for the binary data,

$$Z_{s,y,a,m,\hat{t}} \sim \text{Bin}(1, \phi_{s,y,a,m,\hat{t}}),$$

where  $Z_{s,y,a,m,\hat{t}} = 1$  for  $\text{CPUE}_{s,y,a,m,\hat{t}} > 0$  and  $Z_{s,y,a,m,\hat{t}} = 0$  for  $\text{CPUE}_{s,y,a,m,\hat{t}} = 0$ . The indices  $s$ ,  $y$ ,  $a$ ,  $m$ , and  $\hat{t}$  denote species, year, area, month, and (estimated) target, respectively, and  $\text{logit}(\phi_{s,y,a,m,\hat{t}}) = \xi_s + \xi_{s,y} + \xi_{s,a} + \xi_{s,m} + \xi_{s,\hat{t}}$ . We then fitted the normal model to the log-transformed positive CPUE data:

$$\log(\text{CPUE}_{s,y,a,m,\hat{t}}) \sim N(\eta_{s,y,a,m,\hat{t}}, \tau_s^2),$$

where  $\eta_{s,y,a,m,\hat{t}} = \eta_s + \eta_{s,y} + \eta_{s,a} + \eta_{s,m} + \eta_{s,\hat{t}} + \eta_{s,y,a} + \eta_{s,y,m} + \eta_{s,\hat{t},m} + \eta_{s,\hat{t},a}$  with

$\eta_{s,y,a}$ ,  $\eta_{s,y,m}$ ,  $\eta_{s,\hat{t},m}$ , and  $\eta_{s,\hat{t},a}$  being the interaction terms between year (or target) and area (or month), and  $\tau_s$  the standard deviation (eq. 8). The final yearly trend was calculated from the delta lognormal distribution by area weighting of management strata when the interaction term of year and area was significant (Lo et al. 1992, Maunder and Punt 2004). The confidence intervals of relative abundance trends were approximately calculated assuming a lognormal distribution for relative abundances and using the variance estimates from the Hessian matrix (Buckland et al. 2015). All analyses were conducted using the free statistical software R (R Core Team 2016).

## Results

### Comparison of model performance using simulation tests

The data generated from the simulation tests have linear trends with intercepts that differ depending on the unobserved variable ( $z$ ) and the observed variable ( $e$ ) (Fig. 1). If we knew the unobserved variable, we would be able to infer the yearly trend with a high level of precision and without bias. However, when  $z$  is unknown, the trend is likely to be seriously biased because the model mistakenly evaluates the effect of annual changes in intercepts (i.e., time-varying catchabilities) as a yearly trend. Because we have two species in this simulation, we can use the species composition information instead of the unobserved variable  $z$ . Regression models that use such a surrogate variable are expected to yield an unbiased yearly trend.

When we used 50% or 75% as the critical value ( $X$ ) for Biseau's DC, the biases were much greater than when a critical value of 25% was used. Consequently, in subsequent

analyses, we focused on DC with the 25% critical value. Of course, the complete data model provides an unbiased yearly trend for all scenarios in Table 2 (see Fig. 2, where we focus on the results for species 2 because the effect was more pronounced). The incomplete data model (which ignores the variable  $z$ ) provides highly biased results for all scenarios, even when there is no area shift effect (i.e.,  $\varepsilon = [0, 0]$ ). The degree of bias for the incomplete data model was consistently large for all scenarios. Biseau's DC provided nearly unbiased results for all scenarios except the most complex one (Scenario 4). Although DPC (Winker et al. 2013, 2014) provided better results than the incomplete data model, it provided results with a relatively substantial bias compared with results from DC for scenarios 1 to 3 and those from DRM for all scenarios. CL (He et al. 1997) provided worse results than nZ except in Scenario 4. Importantly, DPC and CL provided a biased result even for the scenario in which the yearly trend is unrelated to  $z$ , whereas all other methods provided unbiased results for this scenario (Fig. 3). DRM provided nearly unbiased results for all scenarios (Figs. 2 and 3). In addition, DRM consistently provided shorter box plots, indicating greater precision, than DC did for all scenarios because the precision of DC decreases due to data subsetting as the heterogeneity of factors associated with the CPUE trend increases. For the most complex scenario (Scenario 4), DC provided almost the same bias as nZ. The reason for this result is that subsetting the data excluded the area effect variable, even when we kept most of the data (75% in this case). Because positive  $e$  decreases when positive  $z$  increases in this scenario (the trends of  $e$  and  $z$  are opposite in Scenario 4 [Table 2]), extracting the data with  $z = 1$  is inconsistent with keeping the area variable  $e$  with sufficient contrast (i.e., both 0 and 1). In contrast, DRM can use all the data and therefore maintains the effect of the area variable without a loss of information. Consequently, DRM provides an almost unbiased

yearly trend. When using the probit or complementary log-log transformation instead of the logit transformation in DRM, the results do not change except that the outliers increase slightly for the complementary log-log transformation (Supplementary Fig. D1).

Taking the probability of exponentiated residuals clarifies the pattern of the unobserved variable, as shown in Fig. 4, which displays the result for one dataset in Scenario 4 (the most complex scenario), except where  $\delta_1 = 1$  and  $\delta_2 = 10$ . If we do not take residuals (Fig. 4a), then the lines of  $(z, e) = (0, 0)$  and  $(z, e) = (1, 1)$  significantly overlap. Because we can observe  $e$  but not  $z$ , discrimination of  $z$  becomes difficult. However, when we take residuals (Fig. 4b), the pattern becomes considerably clearer and easily distinguishable for the combination of  $e$  and  $z$ . In fact, the EM algorithm provided perfect discrimination for the residual data, whereas 476 of 1,000 points in this sample were mistakenly discriminated in the raw data. Thus, taking residuals into account makes the fitting of the mixture model by the EM algorithm easier and more accurate. DPC and CL, which do not consider the residuals, provided lower AIC values when  $\log(\text{CPUE})$  was unrelated to  $z$ , whereas the incomplete data model (= the true model in Scenario 0) and DRM provided the same AIC values (Fig. 3). This biased trend estimate and lower AIC values for DPC were also observed in Scenario 1, whereas DRM provided unbiased trend estimates and the same AIC values as the correct model in Scenarios 0 and 1 (Fig. 3).

When the number of species was increased from two to four, the general trend of results did not change very much (Supplement B). DRM was still the best performer in terms of unbiased trend estimation, although it showed a slight bias for complex scenarios (Scenarios 2 and 3 in Supplement B). DPC analyses with different numbers of PC scores sometimes showed different directions of bias, indicating the importance of selecting the best-performing number of PCs as covariates (Winker et al. 2014).

### Application of the new model to the arabesque greenling data

The number of fishing tactics selected by CAIC was three (the CAIC values for three, two, and one components were 232,668, 242,113, and 251,184, respectively). The annual proportions of the total CPUE were compared for each of the three species against three estimated fishing tactics (Fig. 5). Arabesque greenling was mainly caught by two fishing tactics with different catchabilities (medium gray and light gray in Fig. 5a), whereas one fishing tactic caught few arabesque greenlings over the years (dark gray in Fig. 5a). The two main fishing tactics were relatively stable over the years. One of the main fishing tactics for arabesque greenling did not catch many of the other two species (medium gray region in Fig. 5). One fishing tactic for the variation of CPUE for walleye pollock dominated the others (dark gray in Fig. 5b), indicating that this fishing tactic targeted walleye Pollock, although the fishing tactic also resulted in many Pacific cod being caught (Fig. 5c). Pacific cod were caught primarily by two fishing tactics (dark gray and light gray regions in Fig. 5c). If the tactic accounting for the largest proportion of the total CPUE (dark gray region) corresponds to the tactic mainly targeting walleye pollock, then the second tactic (light gray region) is likely to be mainly targeting Pacific cod. Because the vertical and horizontal habitats of the three species differ, these results suggest that fishers changed their fishing behavior and targeted species based on changes in the catch or abundance of each species.

The AIC for the target-based standardized CPUE using DRM was much smaller than that of the standardized CPUE without the target variable ( $\Delta AIC \approx 40,000$ ). The general trends of standardized CPUEs weighted by the areas of management strata were similar

regardless of the inclusion of the restored target variable and its interactions (Fig. 6). However, when we examine these trends in greater detail, certain differences become apparent (Supplementary Fig. E1). The fittings of non-target CPUE trends were substantially worse than those of the target-based CPUE trends, particularly for target variables 2 (medium gray in Fig. 5) and 3 (dark gray in Fig. 5), which indicates that the predictability of the model without the target variable is much lower. Most of the large differences in AIC are attributable to differences in AIC for binary data. In particular, 93% of the fishing tactics with target variable 3 comprise zero catch. Thus, the hypothesis that fishery efficiency for arabesque greenling depends on fishing tactic seems plausible. Although fishers have recently been encouraged to cut catches and efforts by 30% per year (Morita 2015), a recovery trend indicating the success of such conservation efforts has not been observed, even when taking different fishing tactics into account. Given the dramatic decrease of the arabesque greenling abundance index, additional recovery efforts need to be implemented immediately.

## Discussion

We have viewed the problem of unobserved fishing tactics as a missing data problem. Many multispecies fisheries do not provide information regarding the targeted species in each fishing operation. As is done in traditional multispecies CPUE analyses (Biseau 1998, Stephens and MacCall 2004, Quirijns et al. 2008, Thorson et al. 2016), we used species composition data as a source of information about different fishing tactics. Specifically, the relative proportion of a species in the catch or CPUE being high indicates that this species was targeted in the fishing operation. Differences in fishing tactics can

be considered differences in the intercept terms of regression analyses or differences in catchability. Therefore, we treat this problem as a regression problem with missing data that can be restored by using the species composition data in multispecies fisheries. The key statistical technique for the restoration of missing information is the mixture of regression models for transformed values of proportions of residuals. The posterior probabilities of membership of the components of the mixture correspond to the restored target variables. One advantage of our method is that it does not require subsetting, unlike traditional methods (Biseau 1998, Stephens and MacCall 2004). Keeping all data in the analysis not only improves the precision of the yearly trend estimate but also eliminates or reduces bias by retaining other necessary explanatory variables, as shown in our simulation tests (see also Supplement B). In addition, selection among the models either taking into account or ignoring fishing tactics can be conducted by traditional statistical methods such as AIC, as shown in our simulation tests (Fig. 3) and in the arabesque greenling example.

We used the logit-transformation of the proportion of residuals. The reason for this is that it simplifies our analysis because the transformation permits the use of univariate analysis and a mixture of normal distributions. Because taking the logit transformation results in an approximation, it may sometimes produce an erroneous restoration of the target variables. However, because DRM attributes the problem to the classification of finite mixture components, we expect that restoration will not be highly inaccurate so long as the number of mixture components is well reflected in the residuals. Although it would be ideal to use simultaneous estimation, as in Thorson et al. (2011), it would make the analysis more difficult and complex. Alternatively, we could use vector GLM (Yee 2010) for the multiple residuals instead of the mixture model for finding the pattern from

the proportions of exponentiated residuals. We need to carry out further simulations to determine which method is better.

Mixture models for modelling the error structure of probabilistic distributions, such as delta-lognormal or zero-inflated models, have been extensively applied to CPUE standardization of single-species and multispecies fisheries (Lo et al. 1992, Minami et al. 2007, Thorson et al. 2011, Okamura et al. 2012). In this study, we used a mixture regression model to restore the unobserved target variable from the residual-related information. This is a crucial difference between our model and previous applications of mixture models to CPUE data. After restoring the target variable, more generalized models, such as zero-inflated models (Maunder and Punt 2004), can also be used for CPUE standardization.

Directed CPUE (Biseau 1998) also exhibited good performance in our simulation tests, except in the most complex case. One weakness of this approach is poor precision and accuracy due to the loss of information caused by subsetting. Using the appropriate explanatory variables in directed CPUE analysis is expected to lead to a nearly unbiased yearly trend. However, subsetting occasionally creates difficulties related to the inclusion and selection of appropriate explanatory variables.

Cluster analysis (He et al. 1997) and directed PCA (Winker et al. 2013, 2014) consistently performed worse than DRM in the simulation tests, although DPC generally outperformed the method that ignored the target variable (nZ). Although Winker et al. (2014) proposed an external effective and feasible selection procedure in the form of the optimal coordinate test, we did not use such a procedure in our simulation, and so the performance of DPC may have been unfairly degraded. In addition, AIC did not work well for the scenario in which fishing tactics had no impact when CL and DPC were used



(Fig. 3), which is due to the fact that the information contained in the CL and PC predictor variables from multispecies CPUE composition might not be entirely independent from the response CPUE, as noted in Winker et al. (2014). An additional problem related to this issue is that CL and DPC are unable to drop target-related variables correctly, even when there is no difference in fishing tactics. Because DRM includes a step for selecting the appropriate number of components in the mixture model, it can select the null model that is unrelated to fishing tactics. However, determining the number of clusters or PC scores in advance for CL and DPC does not indicate whether including the target variable in the regression model is necessary, which may produce another problem related to overfitting. Although DRM outperformed CL and DPC for simulations with four stocks and four fishing tactics (Supplement B), one problem of DRM is that mixture regression models are very sensitive to initial values. Therefore, when many species or fishing tactics are present, the computational costs of DRM will increase dramatically, and stable estimation of correct mixture components will become increasingly difficult. In contrast, CL and DPC do not suffer from the problem of selecting appropriate initial values and they may therefore outperform DRM when the number of fishing tactics is high. Reducing the number of species through dimensionality-reduction methods (James et al. 2013) should be useful in such situations. In addition, when the sample size is low (e.g., for count data), the mixture model for normal linear regressions may not work well. In that case, it will be worthwhile to extend our approach to the mixture model for GLMs and their variants (Bishop 2006). In the future, more extensive simulations (e.g., a simulation in which the dataset has many zeros) need to be conducted to further investigate the performance of DRM.

Thorson et al. (2016) used residuals as an effect of the local variation of fishing tactics

similarly to our approach. As noted in Thorson et al. (2016), two-stage modeling techniques like DRM have several drawbacks. Although DRM currently does not include the uncertainty involved in estimating the target variable, incorporation of such uncertainty would be possible with a parametric bootstrap method that assumes the target variable follows a multinomial distribution with the posterior probability of membership of the components in the mixture as the expected values. In addition, because our treatment of residuals includes certain approximations, the performance of one-stage modeling techniques, such as that used in Thorson et al. (2016), are expected to occasionally perform better than DRM. An advantage of our approximation method using two-stage modeling is that it simplifies the analysis and easily permits flexible modeling, such as the inclusion of interaction terms between the target variable and other (observed) variables. Although Thorson et al. (2016) used a continuous variable related to local fishing tactics, we prefer using a discrete variable because it simplifies interpretation. However, maintaining a discrete variable and including interaction terms between the latent target variable and other variables in the one-stage model seems difficult at this stage due to discrete parameters and multivariate non-normal random effects. Nonetheless, we expect that DRM and the spatial dynamic factor analysis will be thoroughly compared using extensive simulation tests and that the two methods will ultimately be integrated into more reasonable one-stage modeling (Thorson et al. 2016) in the future.

We did not assume the existence of an interaction term between year and unobserved target variable because we had no way to determine which of the yearly trends for target and bycatch is representative of the abundance index when those trends are different. In general, one would expect that the target fishery is representative; however, if the bycatch fishery is more survey-like, then the bycatch may be a better indicator. If we had a means

of determining which is better, we could extract a yearly trend on that basis. In that case, DRM would, in principle, be better than traditional methods because DRM directly accounts for the other explanatory variables. In addition, when the targeting effect changes continuously over years, the interaction between the year and the target variable will become significant. For that case, we need to develop a method to extract the year trend of the population.

New predictive techniques developed for statistical machine learning have recently been extensively used for quantitative fishery stock assessment, including CPUE standardization (Kawakita et al. 2005) and global marine ecosystem assessment (Komori et al. 2016). The basic idea of our method should be widely applied to inference problems caused by multispecies data with missing information, not only in fishery science but also in ecological survey data analysis. Unbiased trend estimates and the stock status information obtained thereby can contribute to successful conservation and management. We hope that our approach will become a fundamental tool for CPUE standardization in multispecies fisheries, taking advantage of the efficiency of modern personal computers.

## Acknowledgements

This study was funded by the JST-CREST program. The bottom trawl catch and effort dataset was compiled by the Fisheries Agency and the Fisheries Research and Education Agency of Japan. We greatly thank the associate editor together with H. Winker and one anonymous reviewer for their constructive comments on an earlier draft.

## References

- Biseau, A. 1998. Definition of a directed fishing effort in a mixed species trawl fishery, and its impact on stock assessments. *Aquat. Living Resour.* **11**(3): 119–136. doi:10.1016/S0990-7440(98)80109-5.
- Bishop, C.M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**: 345–370.
- Buckland, S.T., Rexstad, E.A., Marques, T.A., and Oedekoven, C.S. 2015. *Distance Sampling Methods and Applications*. Springer, New York.
- Burnham, K.P., and Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.
- Campbell, R.A. 2016. A new spatial framework incorporating uncertain stock and fleet dynamics for estimating fish abundance. *Fish Fish.* **17**: 56–77.
- Carruthers, T.R., Ahrens, R.N., McAllister, M.K., and Walters, C.J. 2011. Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices. *Fish. Res.* **109**: 157–167.
- Claeskens, G., and Hjort, N.L. 2008. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, UK.

- Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy Stat. Soc. B* **39**: 1–38.
- Geromont, H.F., and Butterworth, D.S. 2015. Generic management procedures for data-poor fisheries: forecasting with few data. *ICES J. Mar. Sci.* **72**: 251–261.
- He, X., Bigelow, K., and Boggs, C. 1997. Cluster analysis of longline sets and fishing strategies within the Hawaii-based fishery. *Fish. Res.* **31**: 147–158.
- Hilborn, R., and Walters, C.J. 1992. *Quantitative Fisheries Stock Assessment: Choice, Dynamics, and Uncertainty*. Chapman and Hall, New York.
- Hokkaido Fisheries Experiment Station 2010. Stock and fisheries of Arabesque greenling (*Pleurogrammus azonus*) in the waters around Hokkaido. Hokkaido Fisheries Experiment Station Technical Report No. 6 (in Japanese)  
[http://www.hro.or.jp/list/fisheries/research/central/kouhou/att/TechnicalReport\\_6.pdf](http://www.hro.or.jp/list/fisheries/research/central/kouhou/att/TechnicalReport_6.pdf)
- James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York.
- Kawakita, M., Minami, M., Eguchi, S., and Lennert-Cody, C.E. 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. *Fish. Res.* **76**: 328–343.

- Komori, O., Eguchi, S., Ikeda, S., Okamura, H., Ichinokawa, M., and Nakayama, S. 2016. An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution* **7**(2), 249-260. doi:10.1111/2041-210X.12473.
- Lo, N.C., Jacobson, L.D., and Squire, J.L. 1992. Indices of relative abundance for fish spotter data based on delta-lognormal models. *Can. J. Fish. Aquat. Sci.* **49**: 2515–2526.
- Lukočienė, O., and Vermunt, J.K. 2010. Determining the number of components in mixture models for hierarchical data. *In* *Advances in Data Analysis, Data Handling and Business Intelligence. Edited by A. Fink, L. Berthold, W. Seidel, and A. Ultsch.* Springer, Berlin-Heidelberg, pp. 241–249.
- Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. *Fish. Res.* **70**(2-3): 141–159.
- McLachlan, G.J., and Krishnan, T. 2008. *The EM Algorithm and Extensions.* Wiley, New Jersey.
- Minami, M., Lennert-Cody, C.E., Gao, W., Román-Verdesoto, M. 2007. Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fish. Res.* **84**(2): 210–221.
- Morita, S.H. 2015. Stock assessment and evaluation for arabesque greenling

- (*Pleurogrammus azonus*). In: Marine fisheries stock assessment and evaluation for Japanese waters. Fisheries Agency and Fisheries Research Agency of Japan, Tokyo. pp 1084–1115. [in Japanese]
- Morita, S.H., Morita, K., Hamatsu, T., Chimura, M., Yamashita, Y., Sasaki, K., Sato, T. 2015. Differential effects of the environment on the growth of arabesque greenling (*Pleurogrammus azonus*): does rising temperature benefit young but not old fish? Environ. Biol. Fish **98**: 801–809
- Narimatsu, Y., Kakehi, S., Ito, S., Okazaki, Y., Inagawa, R., Yano, T. 2015. Impact of the Great East Japan Earthquake tsunami on growth and survival of Pacific cod (*Gadus macrocephalus*) . Can. J. Fish. Aquat. Sci. **72**: 1629–1638.
- Okamura, H., Punt, A. E., and Amano, T. 2012. A generalized model for overdispersed count data. Popul. Ecol. **54**: 467–474.
- Pawitan, Y. 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford, Oxford, UK.
- Quirijns, F.J., Poos, J.J., and Rijnsdorp, A.D. 2008. Standardizing commercial CPUE data in monitoring stock dynamics: Accounting for targeting behavior in mixed fisheries. Fish. Res. **89**(1): 1–8.
- R Core Team. 2016. R: A Language and Environment for Statistical Computing. R

Foundation for Statistical Computing, Vienna. <http://www.r-project.org/>.

Sakurai Y. 2007. An overview of the Oyashio ecosystem. *Deep-Sea Res. II* **54**: 2526–2542.

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

Thorson, J.T., Stewart, I.J., and Punt, A.E. 2011. Accounting for fish shoals in single- and multi-species survey data using mixture distribution models. *Can. J. Fish. Aquat. Sci.* **68**: 1681–1693.

Thorson, J.T., Fonner, R., Haltuch, M.A., Ono, K., and Winker, H. 2016. Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. *Can. J. Fish. Aquat. Sci.* (e-First Article): pp. 1–14.

Tu, C.-Y., Tian, Y., and Hsieh, C.-H. 2015. Effects of climate on temporal variation in the abundance and distribution of the demersal fish assemblage in the Tsushima warm current region of the Japan Sea. *Fish. Oceanogr.* **24(2)**: 177–189.

Walters, C.J. 2003. Folly and fantasy in the analysis of spatial catch-rate data. *Can. J. Fish. Aquat. Sci.* **60**: 1433–1436.

Walters, C.J., and Martell, S.J.D. 2004. *Fisheries Ecology and Management*. Princeton,



Princeton, USA.

Winker, H., Kerwath, S.E., and Attwood, C.G. 2013. Comparison of two approaches to standardize catch-per-unit-effort for targeting behaviour in a multispecies hand-line fishery. *Fish. Res.* **139**: 118–131.

Winker, H., Kerwath, S.E., and Attwood, C.G. 2014. Proof of concept for a novel procedure to standardize multispecies catch and effort data. *Fish. Res.* **155**: 149–159.

Yee, T. W. 2010. VGLMs and VGAMs: An overview for applications in fisheries research. *Fish. Res.* **101**: 116–126.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York, USA.

Table 1. The effect of using residuals. See the text for the parameters used for population changes.

		Raw log(CPUE)		Residuals of log(CPUE)	
	<i>q</i>	Year 1	Year 2	Year 1	Year 2
Target =1					
Species 1	0.10	4.61	3.69	0.35	0.35
Species 2	0.01	2.30	0.69	−0.35	−0.35
Target = 2					
Species 1	0.05	3.91	3.00	−0.35	−0.35
Species 2	0.02	3.00	1.39	0.35	0.35

Draft

Table 2. Control parameters used for four scenarios in the simulations:  $\delta_s$  is the scale parameter for species  $s$  ( $s = 1, 2$ );  $\gamma_s$  is the intercept (or catchability) that changes depending on the unobserved variable  $z$ ;  $\varepsilon_s$  is the intercept that changes depending on the observed variable  $e$ ;  $\beta_s$  is the true yearly trend for species 1 and 2;  $a$  and  $b$  are the intercept and slope for the yearly transition of  $z$ , respectively;  $c$  and  $d$  are the intercept and slope for the yearly transition of  $e$ , respectively;  $\sigma_s$  is the standard deviation that corresponds to the coefficient of variation in the exponentiated scale.

Scenario	$(\delta_1, \delta_2)$	$(\gamma_1, \gamma_2)$	$(\varepsilon_1, \varepsilon_2)$	$(\beta_1, \beta_2)$	$(a, b)$	$(c, d)$	$(\sigma_1, \sigma_2)$	Description
1	(1, 2)	(10, 2)	(0, 0)	(-1.5, -0.5)	(1, -0.2)	(1, -0.2)	(0.8, 0.8)	No areal effect. The yearly transition of $z$ is mild.
2	(1, 2)	(10, 2)	(0.5, 0.5)	(-1.5, -0.5)	(1, -0.2)	(1, -0.2)	(0.8, 0.8)	The yearly transitions of $z$ and $e$ are mild. No interaction between $z$ and $e$ .
3	(1, 2)	(10, 2)	(0.1, 0.9)	(-1.5, -0.5)	(2, -0.5)	(2, -0.5)	(0.8, 0.8)	The yearly transitions of $z$ and $e$ are severe. Substantial interaction between $z$ and $e$ .
4	(1, 2)	(10, 2)	(0.1, 0.9)	(-1.5, -0.5)	(2, -0.5)	(-2, 0.5)	(0.8, 0.8)	The yearly transitions of $z$ and $e$ are severe and in opposing directions. Substantial interaction between $z$ and $e$ .

## Figure captions

Fig. 1

The simulation data for each scenario. The bar plots show yearly changes of the target variable  $z$  and the areal variable  $e$ . Embedded subplots show the expected trends for each species in the western and eastern areas in which intercepts depend on the target variable  $z$ .

Fig. 2.

Comparison of yearly trend estimates of species 2 for simulated data.  $Z$ ,  $nZ$ , DC, DPC, CL, and DRM denote the complete data model, the incomplete data model, directed CPUE, directed PCA, cluster analysis, and the directed residual mixture model (target-based CPUE), respectively. As a visual aid, the results of the comparison among  $nZ$ , DPC, and CL are shown in the small interior boxes. The broken horizontal lines correspond to the true value ( $-0.50$ ).

Fig. 3.

Left panels: Trend estimates for species 2. Right panels: AIC values for species 2.  $Z$ ,  $nZ$ , DRM, DPC, and CL denote the complete data model, the incomplete data model, the directed residual mixture model, directed PCA, and cluster analysis, respectively. Scenario 0 (SC0) is the additional scenario where the response variable is unrelated to the target variable and the parameters are the same as in Scenario 1 (SC1) except where  $\gamma_1 = \gamma_2 = 2$  (when  $\gamma_1 = \gamma_2$ , the response variable is unrelated to the target variable). SC1 is Scenario 1 (see Table 2).

Fig. 4.

The logit-transformed probability of relative CPUE for a dataset simulated under Scenario 4, except where  $\delta_1 = 1$  and  $\delta_2 = 10$ . The probability calculated (a) for the raw CPUE data  $[\text{logit}(q_1)]$ , and (b) for the residuals of the CPUE data  $[\text{logit}(p_1)]$ .

Fig. 5.

The variation of annual CPUE compositions for three species classified by the restored target variable,  $\lambda_{s,y,\hat{t}} = \frac{CPUE_{s,y,\hat{t}}}{\sum_{\hat{t}=1}^3 CPUE_{s,y,\hat{t}}}$  ( $s = 1, 2, 3; y = 1997, \dots, 2015; \hat{t}$  is estimated from eq. 7). Common color types (dark gray, medium gray, and light gray) are used across the different species indicate the same target variable ( $\hat{t} \in \{1, 2, 3\}$ ). The dotted lines denote the trajectories of mean CPUE for each species.

Fig. 6.

The yearly trends extracted from the standardized CPUEs weighted by area. The broken and solid lines denote the standardization with and without the target variable and its interactions, respectively. The shaded regions in the embedded subplots show approximate 95% confidence intervals.

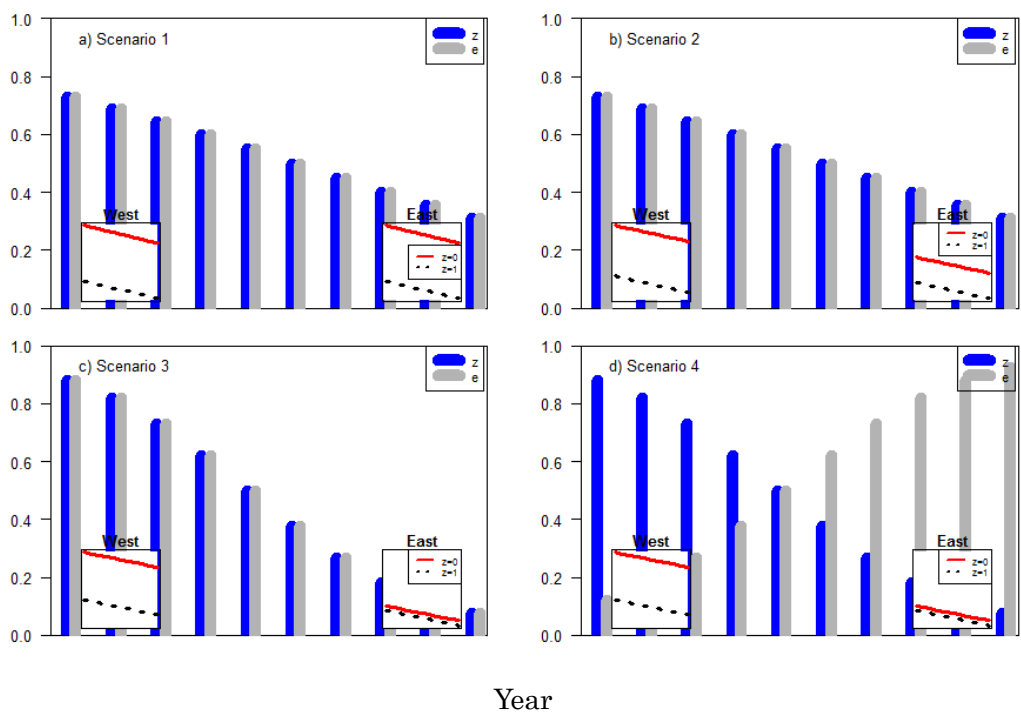


Fig. 1.

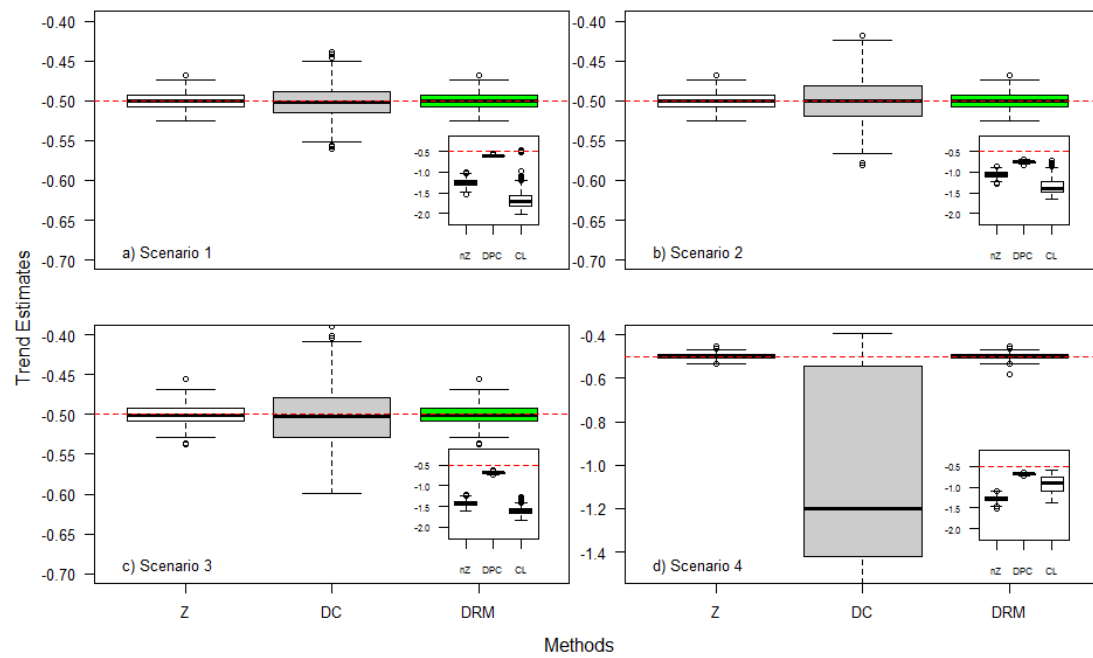


Fig. 2.

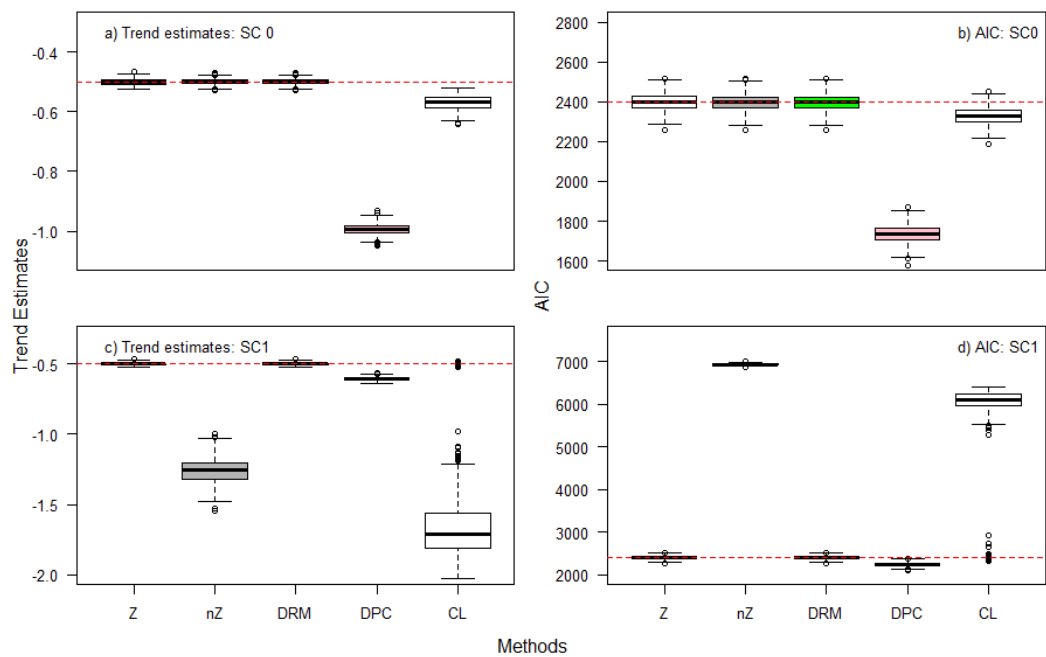


Fig. 3.



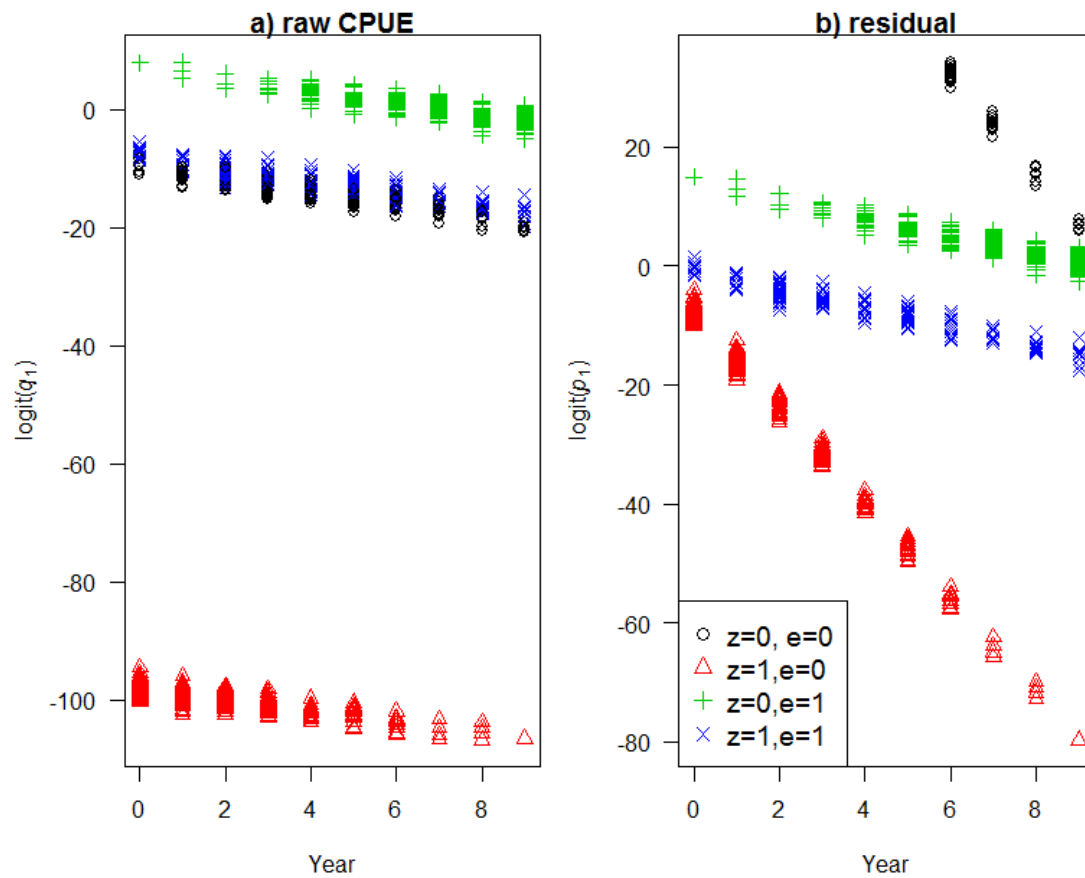


Fig. 4.

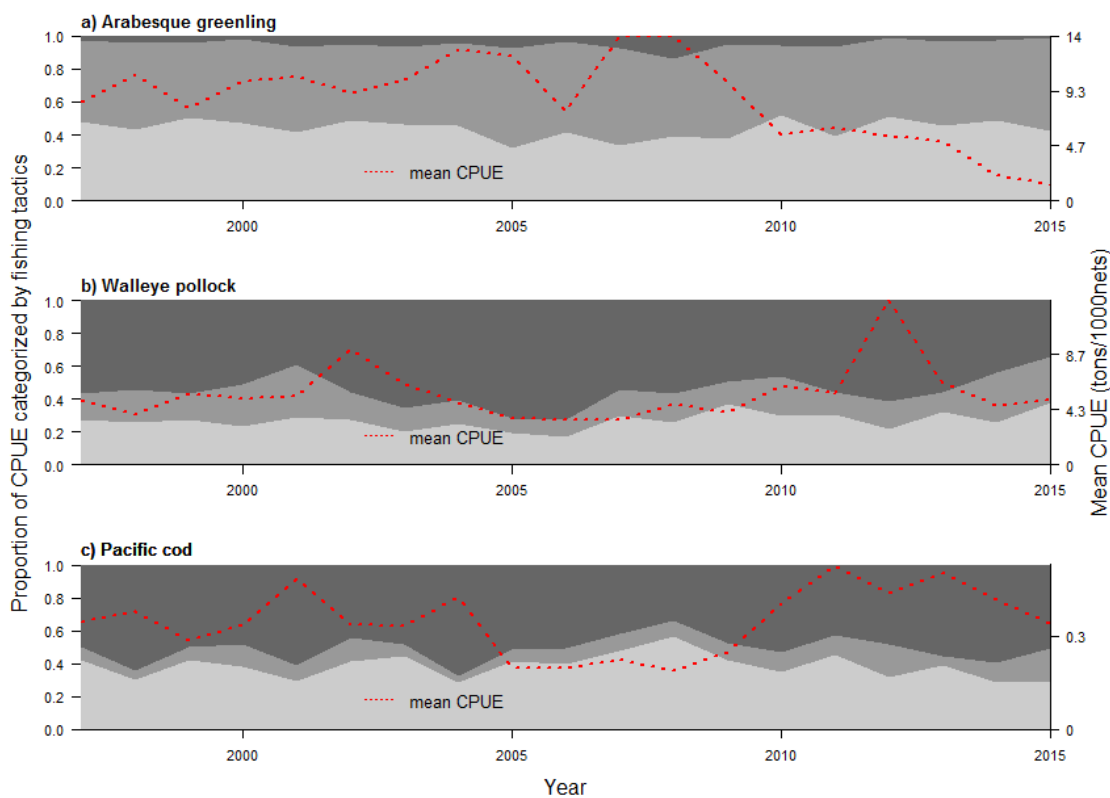


Fig. 5.

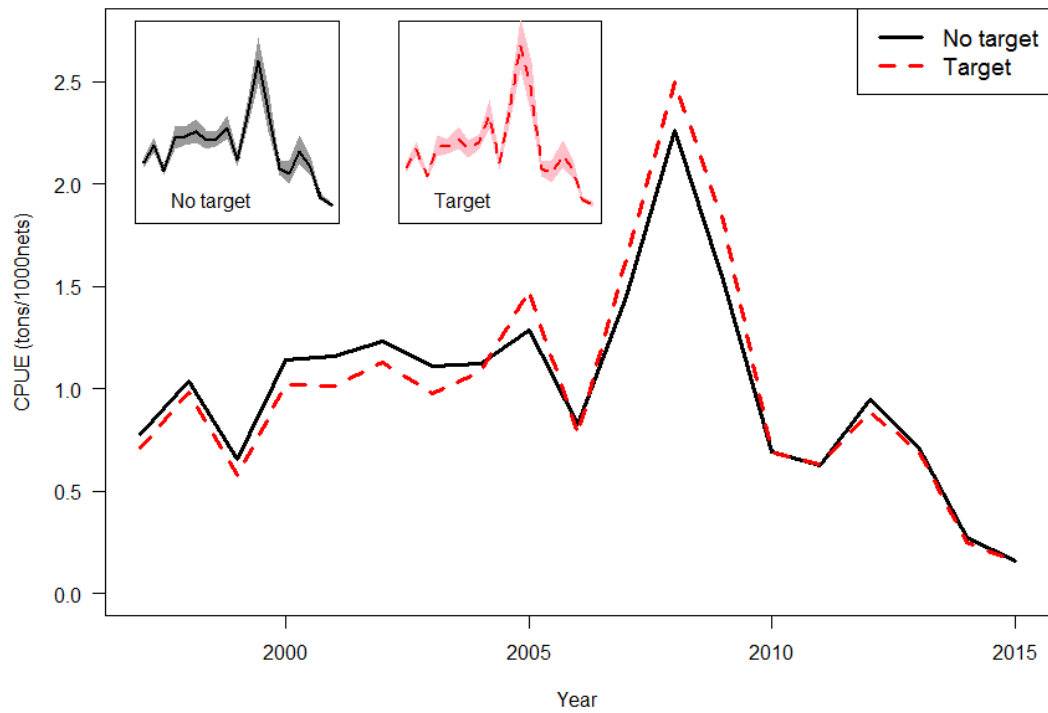


Fig. 6.