Western and
Central Pacific
Fisheries
Commission

SCIENTIFIC COMMITTEE
SIXTH REGULAR SESSION

10-19 August 2010
Nuku'alofa, Tonga

Application of the Tweedie distribution to zero-catch data in CPUE analysis

WCPFC-SC6-2010/ME-WP-02

Hiroshi SHONO [1]

[1] National Research Institute of Far Seas Fisheries, Fisheries Research Agency, 5-7-1, Orido, Shimizu-ku, Shizuoka-shi, Shizuoka-ken 424-8633, Japan

Fisheries Research 93 (2008) 154–162

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Application of the Tweedie distribution to zero-catch data in CPUE analysis

Hiroshi Shono*

National Research Institute of Far Seas Fisheries, Fisheries Research Agency, 5-7-1, Orido, Shimizu-ku, Shizuoka-shi, Shizuoka-ken 424-8633, Japan

ARTICLE INFO

Article history:
Received 8 June 2007
Received in revised form 27 March 2008
Accepted 27 March 2008

Keywords:
By-catch species
CPUE standardization
Generalized linear model
n-fold cross-validation
Tweedie distribution model
Zero-catch problem

ABSTRACT

We focus on the zero-catch problem of CPUE (catch per unit effort) standardization. Because the traditional CPUE model with a log-normal error structure cannot be applied in this case, three methods have often been utilized as follows:

(1) Ad hoc method adds a small constant value to all response variables.
(2) Catch model with a Poisson or negative-binomial (NB) error structure.
(3) Delta-type two-step method such as the delta-normal model (after estimating the ratio of zero-catch using
a logit or probit model, a model such as CPUE log-normal or Catch-Poisson is applied to CPUE without zero-data).

However, there are some statistical problems with each of these methods.

In this paper, we carried out the CPUE standardization mainly using the Tweedie distribution model based on the actual by-catch data (silky shark, *Carcharhimus falciformis*, in the North Pacific Ocean caught by Japanese training vessels) including many observations with zero-catch (>2/3rd) and tuna fishery data as a target (yellowfin tuna, *Thunnus albacares*, in the Indian Ocean caught by Japanese commercial vessels) where the ratio of zero-catch is not so high (<1/3rd). The Tweedie model is an extension of compound Poisson model derived from the stochastic process where the weight of the counted objects (i.e., number of fish) has a gamma distribution and has an advantage of handling the zero-catch data in a unified way.

We also compared four candidate models, the Catch-NB model, ad hoc method, Delta-lognormal model (delta-type two-step method) and Tweedie distribution, through CPUE analyses of actual fishery data in terms of the statistical performance. Square error and Pearson's correlation coefficient were calculated based on the observed CPUE and the corresponding predicted CPUE using the *n*-fold cross-validation.

As a result, the differences in the trend of CPUE between years and model performance between the ad hoc method and Tweedie model were found to be not so large in the example of yellowfin tuna (target species). However, the statistical performance of Tweedie distribution is rather better than Delta-lognormal model, the Catch-NB distribution and ad hoc method in the example of silky shark (by-catch species). Standardized CPUE year trend of ad hoc method was found to be quite different from that of the Tweedie distribution and other two models. Model performance of the Tweedie distribution is good judging from the 5-fold cross-validation using the fishery data if including many zero-catch data such as by-catch species.

1. Introduction

Catch per unit effort (CPUE) is an important concept which is corresponding to the relative stock size and usually assumed to be proportional to the stock abundance. However, the nominal CPUE may include spatiotemporal effects such as area, season and various environmental factors like sea surface temperature. In order to remove these various effects from the nominal CPUE and to extract the year trend proportional to the stock density, some "standardization" has often been performed using various statistical methods. Such methods are called CPUE standardization (Gavaris, 1980).

CPUE standardization is nowadays indispensable for fish stock assessment because standardized CPUE is used for stock assessment models as a tuning index and it has largely affects on the estimated results of stock status in many cases.

As a statistical method, the generalized linear model (GLM) (Dobson, 1990) including analysis of covariance (ANCOVA), which

* Tel.: +81 543 36 6000x43; fax: +81 543 35 9642.
E-mail address: hshono@affrc.go.jp.

0165-7836/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2008.03.006

is expressed as the following formulae in the matrix form, has been usually used for CPUE standardization.

$$Y = X\Theta + \varepsilon,$$
$$Y = (y_i) = \begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix}, \qquad X = (x_{ij}) = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{s1} & \cdots & x_{sp} \end{bmatrix},$$
$$\Theta = (\theta_j) = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}, \qquad \varepsilon = (\varepsilon_i) = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_s \end{bmatrix}$$
$$E[Y] = \mu = g^{-1}(X\Theta),$$
$$\mathrm{Var}[Y] = \phi\,\mathrm{Var}[\mu]$$

where Y: response variable; X: explanatory variables (covariate vector/design matrix); Θ: unknown parameters; ε: error term; g( ): link function; ϕ: dispersion parameter; μ: expectation of the response variable; suffix $i = 1, \ldots, s$ (s: sample size) and $j = 1, \ldots, p$ (p: number of parameters).

In this study, we mainly applied the two types of GLM, CPUE model (model-A) and catch model (model-B), which are described as follows:

$$E[Y] = E[\mathrm{CPUE}] \quad \text{or} \quad E[\log(\mathrm{CPUE})] = \mu = \mathrm{g}^{-1}(X\Theta) \quad \text{(model-A)}$$

$$E[Y] = E[\mathrm{Catch}] = (\mathrm{Effort}) \times \mu = (\mathrm{Effort}) \times \mathrm{g}^{-1}(X\Theta) \quad \text{(model-B)}$$

Remark. (Effort) is basically assumed as the offset in the model-B.

In many CPUE analyses, the following CPUE log-normal model (i.e., CPUE model with log-normal error, ANCOVA-type) is often utilized.

$$\begin{aligned} E[\log(\mathrm{CPUE})] = {} & (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Area}) + (\mathrm{Season}) \\ & + (\text{enviornmental factors, fishing gears,} \\ & \text{operating devices, etc.}) + \cdots + (\mathrm{Interactions}) \end{aligned} \quad (1)$$

where (Year): effect of year, (Area): effect of area; (Season): effect of month/quarter; (environmental factors, fishing gears, operating devices, etc.): effect of environmental factors such as SST, fishing gears, operating devices, etc. (Interactions): two way interactions, $\log(\mathrm{CPUE}) \sim N(\mu, \sigma^2)$ and g( ): identity.

However, the analysis of a covariance model with a normal error structure cannot be applied to the “zero-catch” data with which catch is equal to zero because the natural logarithm of zero is equal to negative infinity. Several statistical approaches have been previously used for the zero-catch problem in the field of fish stock analysis as follows:

(1) Use of the ad hoc method that a constant (viz. a small value) is added to all response variables (i.e., CPUE) as follows (Robson, 1966):

$$\begin{aligned} E[\log(\mathrm{CPUE} + \mathrm{constant})] = {} & (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Area}) \\ & + (\mathrm{Season}) + (\mathrm{EMT}) + \cdots \\ & + (\mathrm{Interactions}) \end{aligned} \quad (2)$$

(2) Use of the Catch-Poisson or Catch-Negative-Binomial (NB) regression model (i.e., Catch model with Poisson/negative-binomial error, GLM-type). In these models, catch not CPUE, which is defined as a categorical variable, is set to the response variable (Reed, 1996) in the model-B described previously.

(3) Use of the delta-type two-step model (e.g., Delta-lognormal model) (Lo et al., 1992) or zero-inflated model (Lambert, 1992).

In this two-step method, the ratio of zero-catch is estimated by the logit model with logit-link function in the 1st step as:

$$\begin{aligned} \log\left(\frac{q}{1-q}\right) = {} & (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Area}) + \cdots + (\mathrm{Interactions}) \\ & + (\log(\mathrm{Effort})) \end{aligned} \quad (3)$$

where $E[X] = q$, $X \sim \text{Binomial}(\theta)$, $g(x) = \log(x/(1-x))$ and

$$X = \begin{cases} 1 & (\text{if Catch} > 0) \\ 0 & (\text{Otherwise}) \end{cases}.$$

After that, the CPUE log-normal or Catch models to the data with a positive catch (i.e., CPUE) is applied in the 2nd step.

In fact, the ad hoc method (1) (that the small constant value is uniformly added to all response variables) has been previously often used. Although this way is easy to carry out, it leads to a bias in the interval estimate. Theoretically it is possible to avoid this bias by subtracting the constant from the estimate with regards to the point estimation. Although the estimation of unknown parameters is actually performed by adding a small value to all CPUE, in practice this incurs a bias. We also have a problem in the ad hoc method regarding what value is adequate as a constant term to add to all response variables. Although ICCAT (International Commission for the Conservation of Atlantic Tunas) recommended the use of 10% of the overall mean CPUE as the constant (Anon., 1997), the reason seems not to be clear.

Catch model with Poisson or negative-binomial error structure (2) has been utilized for CPUE standardization in the late 1990s. Although the Catch-Poisson model was initially used, the constraint that the expectation is equivalent to the variance is too strict (i.e., the Catch-Poisson model does not fit observed CPUE well). Thus, the Catch-negative-binomial model has been gradually applied instead of Poisson distribution after the negative binomial distribution was included into the GLM procedure of SAS/STAT package (Version 9.1; SAS Institute Inc., Cary, NC, USA) (SAS, 2004).

Delta-type two-step model (3) that (i) estimates the ratio of zero-catch by logit or probit model in Eq. (3) and (ii) applies the CPUE-log-normal or Catch model to the part of non-zero data has been recently used for CPUE standardization. In this two-step model, the combination of explanatory variables which are statistically significant may differ in step (i) and (ii), and this complicates model interpretation. In addition, it is sometimes difficult to include the (Year × Area) interactions as a fixed effect especially in the 1st step of the delta-type two-step model due to the missing data even though which generally appears to be statistically significant because of the wide range of the spatiotemporal movement in the tuna species. For general discussion about various problems of CPUE standardization including the issue of the (Year × Area) interaction, see Maunder and Punt, 2004.

In the delta-type two-step model, it is possible to estimate unknown parameters in step (i) and (ii) simultaneously by connecting log-likelihood functions and this method is called the zero-inflated model (Lambert, 1992). Zero-inflated Poisson/negative-binomial model was applied for CPUE standardization of by-catch species caught by the purse seine fishery (Kawakita et al., 2005). Because the zero-inflated models are not included into the exponential family different from the Tweedie distribution, there is a possibility of misspecification of the structure in the zero-inflated model as follows:

- The consistent estimator of the maximum likelihood estimator may not be obtained under only the assumption of the variance structure.
- The moment estimator may not be asymptotic equivalent to the maximum likelihood estimator and so on.

In fact, the exponential family with equal to or less than quadratic variance function is limited to the following six probability distributions (Normal, Gamma, Poisson, Hyperbolic secant, binomial and negative binomial) (Morris, 1982).

In this study, we discuss the zero-catch problem, which means that CPUE-lognormal model (where the natural logarithm of CPUE is set as the response variable) cannot be mathematically applied in the case of including zero-catch data. We utilized the Tweedie distribution model, in which zero-data are uniformly dealt with, and compared the Tweedie model to the previously used methods such as the ad hoc method, Catch-NB model and delta-type two-step method (Delta-lognormal model) through two case studies based on the actual fishery catch and effort data. The Tweedie model was applied for CPUE analyses of Patagonian toothfish (Candy, 2004).

The section composition of this paper is as follows.

In this section, we describe the background of zero-catch problem and introduce several ways utilized for CPUE standardization in the case of including zero-catch data.

In the second section, we describe the Tweedie distribution model, mainly the characteristics, advantages and disadvantages, and the procedure of parameter estimation for the following case studies.

In the third section, we compare the Tweedie model with the ad hoc method using the catch and effort data for yellowfin tuna in the Indian Ocean.

In the fourth section, we compare the four models (i.e., Tweedie model, Catch-NB model, Delta-lognormal model and ad hoc method) using the catch and effort data for silky shark in the North Pacific Ocean.

2. Materials and methods

In this section, we focus on the Tweedie distribution model in which zero-data can be uniformly handled (Tweedie, 1984). The probability density function of the Tweedie distribution is expressed in the following formula (4).

$$f(y:\mu,\sigma^2,p)=a(y:\sigma^2,p)\,\exp\left\{-\frac{1}{2\sigma^2}d(y:\mu,p)\right\} \tag{4}$$

where μ: location parameter; σ^2: diffusion parameter; p: power parameter.

Remark. $d(y$: $\mu, p)$ is called the unit deviance.

The Tweedie model can be transformed to the exponential family using an appropriate change of variables. This enables us to discuss using the framework of quasi-likelihood in the generalized linear model (GLM). It is known that these estimates of regression coefficients by the Tweedie model have asymptotically good performance in which the sample mean expresses the likelihood estimator of expectation (Jorgensen, 1997). We can transform into an exponential family though adequate change of variables and the power parameter (p) is also shown in the variance function of the following expression (5). In this paper, we used the following GLM framework shown in Eq. (5) with log-link function as a Tweedie regression model.

$$\begin{aligned}&E[Y](=E[\text{CPUE}]=\mu)\\&\quad=\exp\{(\text{Intercept})+(\text{Year})+\cdots+(\text{Interactions})\}\\&\text{Var}[Y](=\phi\text{Var}[\mu])=\sigma^2\mu^p\end{aligned} \tag{5}$$

where σ^2 is defined by the parameter ϕ in the model-A described previously.

The Tweedie model can express the Poisson, Gamma and inverse Gaussian distributions if the power-parameter (p) is 1, 2, and 3, respectively.

This power-parameter (p) can be defined as an arbitrary real number except for $0<p<1$ and we are mainly interested in the range of $1<p<2$. Because the Tweedie model is expressed as the compound Poisson distribution (described as the formula (6)) if $1<p<2$ and then which seems to be appropriate for CPUE analysis of by-catch species such as shark with a lot of zero-catch data, we mainly focused on such case (if $1<p<2$) in this paper.

$$Y=\left\{\begin{matrix}\displaystyle\sum_{k=1}^{N}X_k & (N=1,2,3,\ldots)\\ 0 & (N=0)\end{matrix}\right\} \tag{6}$$

where $X_1,\ldots,X_N$ is identical and independently Gamma distributed with mean m and variance $\alpha^{-1}m^p$ (α is defined in Eq. (9)) and N is Poisson distributed with mean λ. Variable Y in formula (6) corresponds to the CPUE value in the following case studies. In addition, the probability density distribution of the Tweedie model in Eq. (4) and the (quasi-)deviance (D) are explicitly shown as the following formula (7) if and only if $1<p<2$.

$$\begin{aligned}&d(y:\mu,p)=2\left\{\frac{\max(y,0)^{2-p}}{(1-p)(2-p)}-\frac{y\mu^{1-p}}{1-p}+\frac{\mu^{2-p}}{2-p}\right\},\\&a(y:\sigma^2,\mu)=\left\{\frac{\sigma^{2(\alpha+1)}y^{\alpha}}{(1-p)^{\alpha}(2-p)}\right\}^{n}\frac{1}{n!\,\Gamma(n\alpha)y},\\&D=2\sum_{i=1}^{s}\frac{y_i^{2-p}-(2-p)y_i\mu_i^{1-p}+(1-p)\mu_i^{2-p}}{(1-p)(2-p)},\end{aligned} \tag{7}$$

where $\alpha=(2-p)/(p-1)$ and s is the sample size.

In this study, we carried out the parameter estimation based on the following procedure.

(i) Estimate the power parameter (p) by maximizing the profile log-likelihood across the grid values of (p) (see Figs. 2 and 6) in the range of $1<p<2$ through the explicit form of the probability density function in Eq. (7).

(ii) Estimate the regression coefficients using the framework of quasi-likelihood in the GLM fixing the value of p in the estimate obtained in the step (i) also based on the formulae of Eq. (7) including the deviance.

The probability density function f of the Tweedie distribution in Eq. (4) is also written as follows (Smyth, 1996):

$$\begin{aligned}f(y:\mu,\sigma^2,p)&=P(N=0)d_0(y)+\sum_{l=1}^{\infty}P(N=1)g_{Z|N=1}(y)\\&=\mathrm{e}^{-\lambda}d_0(y)+\sum_{l=1}^{\infty}\frac{\lambda^l\,\mathrm{e}^{-\lambda}}{l!}\,\frac{y^{l\alpha-1}\,\mathrm{e}^{-y/\tau}}{\tau^{l\alpha}\Gamma(l\alpha)}\end{aligned} \tag{8}$$

where d_0 is the Direct delta function, $g_{Z|N}$ is the conditional density of Z given N and Z has the same distribution as Y (shown in Eq. (6)) with

$$\lambda=\frac{\mu^{2-p}}{\sigma^2(2-p)},\qquad \alpha=\frac{2-p}{p-1},\qquad \tau=\sigma^2(p-1)\mu^{p-1},$$
$$m=\sigma^2(2-p)\mu^{p-1} \tag{9}$$

$\mathrm{e}^{-\lambda}$ shows the predicted probability of a zero-catch.

On the other hand, there are some disadvantages that (1) overall model comparison between Tweedie distribution and other statistical models is generally difficult except for the evaluation of each estimate value. (2) Common information criterion such as AIC is not available due to using the framework of quasi-likelihood.

Although a new information criterion Q-AIC (quasi-AIC), which can be applied to the Tweedie distribution model, as suggested by Burnham and Anderson (1998), the theoretical validity seems to be questionable. The Tweedie model becomes an expansion of the power parameter of relationship between mean and variance to continuous variable.

We evaluated the accuracy of the four models using *n*-fold cross-validation that (1) divide all data from the nth sub-datasets randomly (2) calculate the predicted values concealing the observed ones of each sub-dataset on purpose. We utilized the correlation coefficient and the square error between the observed and the corresponding predicted values for validation in the candidate models and checked the trends of standard residual based on the correlation plots.

In this study, we compared the Tweedie model, the ad hoc method (the Catch-negative-binomial model and the Delta log-normal model) using two case studies of actual fishery data, yellowfin tuna in the Indian Ocean by Japanese commercial longline vessels and silky shark in the North Pacific by Japanese training longline vessels. We utilized R (Version 2.5.0) and SAS (Version 9.1.3) for these computations.

2.1. Case study 1: yellowfin tuna in the Indian Ocean by Japanese commercial longline vessels

We performed CPUE standardization for yellowfin tuna in the Indian Ocean using catch and effort data caught by Japanese commercial longline vessels. The purpose is to compare between the ad hoc method where a constant term (viz. small value) is added to all response variables (i.e., CPUE) and the Tweedie distribution model. We used aggregated data (5 × 5 degree square/monthly basis) and the following explanatory variables and response variable. The ratio of zero-catch data is approximately 10%.

Response variable

CPUE (catch in number per 1000 hooks) for yellowfin tuna in the Indian Ocean caught by Japanese commercial longline vessels.

Explanatory variables

Year (1960–2003), Month (1–12), Area (1–5, Fig. 1), Gear (number of hooks between float, HBF), SST (sea surface temperature), MLD (mixed layer depth).

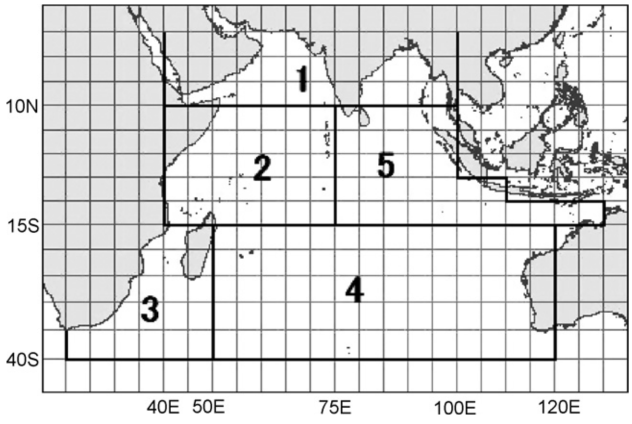

Fig. 1. Area stratification used for CPUE standardization of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial fishery.

Remark. Year, Month, Area and SST, MLD are defined as categorical and continuous variables, respectively. Agreed area stratification for yellowfin tuna in the IOTC working party on the tropical tuna is shown in Fig. 1.

At first, after 0.1 was uniformly added as the constant in the ad hoc method, we selected the final model by BIC (Baysian information criterion: Schwarz, 1978) of the candidate models in the range of null model in Eq. (10) where the main effect of year is only included, to the full model in Eq. (11) in which all main effects and two-way interactions are included. Because the main objective of CPUE analyses is to extract year trend of relative abundance, the main effect of year is included into the null model.

$$\log(\text{CPUE}_i + 0.1) = \text{Intercept} + \text{Year}_i + \text{Error}_i, \quad \text{Error}_i \sim N(0, \sigma^2) \tag{10}$$

$$\begin{aligned}\log(\text{CPUE}_{ijk} + 0.1) &= \text{Intercept} + \text{Year}_i + \text{Month}_j + \text{Area}_k + \text{GEAR} + \text{SST} + \text{MLD}\\ &\quad + (\text{Year} * \text{Month})_{ij} + (\text{Year} * \text{Area})_{ik} + (\text{Month} * \text{Area})_{jk}\\ &\quad + (\text{Year} * \text{SST})_i + (\text{Year} * \text{MLD})_i + (\text{Area} * \text{GEAR})_k + (\text{Area} * \text{SST})_k\\ &\quad + (\text{Area} * \text{MLD})_k + (\text{Month} * \text{GEAR})_j + (\text{Month} * \text{SST})_j\\ &\quad + (\text{Month} * \text{MLD})_j + (\text{SST} * \text{MLD}) + \text{Error}_{ijk}, \quad \text{Error}_{ijk} \sim N(0, \sigma^2)\end{aligned} \tag{11}$$

After estimating the power parameter of variance function by maximizing the profile log-likelihood in the BIC-best model, we estimated each parameter of regression coefficients using the quasi-likelihood framework in the Tweedie distribution model.

2.2. Case study 2: silky shark in the North Pacific Ocean by Japanese training longline vessels

We carried out CPUE standardization for silky shark in the North Pacific Ocean utilizing catch and effort data caught by Japanese training longline vessels. The purpose is to compare the ad hoc method, the Catch-NB (negative-binomial) model (i.e., catch model with negative-binomial error structure), delta-type two-step method (Delta-lognormal model) and the Tweedie distribution model. We used the data for each operation, the following explanatory variables and response variable for this analysis. The ratio of zero-catch is more than 80%.

Response variable

CPUE (catch in number per 1000 hooks) for silky shark in the North Pacific Ocean caught by Japanese training longline vessels.

Explanatory variable

Year (1992–2003), Quarter (1–4, 1: January–March, 2: April–June, 3: July–September, 4: October–December), Area (1–4, 1: $0 \leqq \text{Lat.} < 20$, 2: $20 \leqq \text{Lat.} < 30$, 3: $30 \leqq \text{Lat.} < 40$, 2: $40 \leqq \text{Lat.} < 50$), Gear (number of hooks between float, HBF), Year, Quarter, Area and Gear are set as categorical and continuous variables, respectively.

3. Results

We describe the following results in terms of comparison among the Tweedie distribution model, ad hoc method, Catch-NB model and Delta-lognormal model.

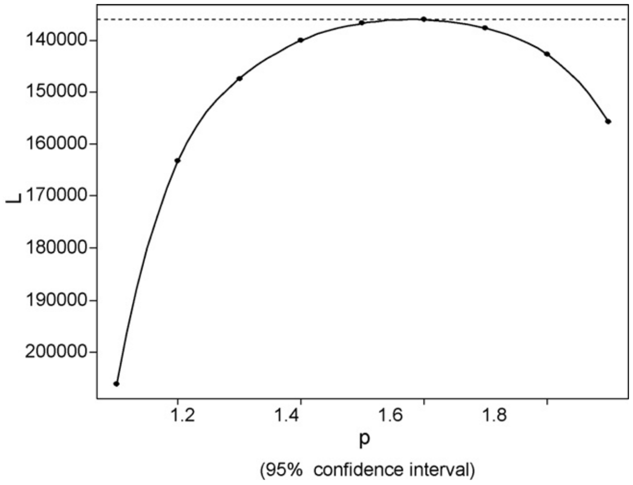

Fig. 2. Value of log-likelihood function (L) changing the power-parameter (p) of the Tweedie model for CPUE standardization of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial fishery.

- Estimation of power-parameter of variance function in the Tweedie model.
- Correlation plots of predicted and observed values by 5-fold cross-validation.
- Residual analyses.
- Extracted CPUE year trends.
- Model comparison by correlation coefficient and square error.

3.1. *Case study 1: Yellowfin tuna in the Indian Ocean by Japanese commercial longline vessels*

In this case study, we estimated the unknown regression parameters of the Tweedie distribution model using the same combination of explanatory factors as the final model in formula (12) obtained from the ad hoc method.

$$\begin{aligned}\log(\text{CPUE}+0.1) = {} & \text{Intercept}+\text{Year}+\text{Month}+\text{Area}+\text{Gear}+\text{SST}\\ & +\text{MLD}+\text{Area}*\text{MLD}+\text{Error},\\ & \text{Error}\sim N(0,\sigma^2)(\text{Final Model}) \qquad (12)\end{aligned}$$

In Fig. 2, X axis and Y show the power parameter (p) and the value of the log-likelihood function. The value of (p) corresponding to the MLE (maximum likelihood estimates) was approximately estimated at 1.58 in this case. Assuming this value (1.58) of power parameter (p), other parameters including regression coefficients were estimated using the framework of the quasi-likelihood method. We also utilized the same combination of explanatory variables in the formula (10) in the Tweedie model as the ad hoc method. Fig. 3 shows the quantile–quantile (QQ) plots based on the value of deviance and standard residual in the Tweedie model and the ad hoc method. The residual skewed a little on the left-hand side because some predicted values obtained from the model corresponding to the zero-catch observations became positive. The difference of the year trends of standardized CPUE between based on the Tweedie model and the ad hoc method, shown in Fig. 4, is not so large although the CPUE year trend obtained from the Tweedie model looks like more stable than that by the ad hoc method.

Next, we describe the accuracy comparison of the Tweedie model and the ad hoc method using 5-fold cross-validation based on the suggestion by Breiman et al. (1984), where they described that it is empirically good in general to set the division number (n) to 5 in the n-fold cross-validation. Table 1 shows the sub-datasets divided randomly. In Table 1, “Rule” and “C.V.” express the supervised data used for parameter estimation of both models and unsupervised ones for cross-validation with concealing the observed values deliberately.

Table 1
Dataset used for 5-fold cross-validation in the example of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial fishery

Sub-set	No. of data	Scenarios					
		Base case	I	II	III	IV	V
1	9871	Rule	C.V.	Rule	Rule	Rule	Rule
2	9872	Rule	Rule	C.V.	Rule	Rule	Rule
3	9871	Rule	Rule	Rule	C.V.	Rule	Rule
4	9872	Rule	Rule	Rule	Rule	C.V.	Rule
5	9871	Rule	Rule	Rule	Rule	Rule	C.V.

Rule and C.V. show the sub-dataset for rulemaking and cross-validation, respectively.

Table 2 shows the overall values of Pearson’s correlation coefficient and square error between the observed and the predicted CPUE values. Values of correlation and square error in each case

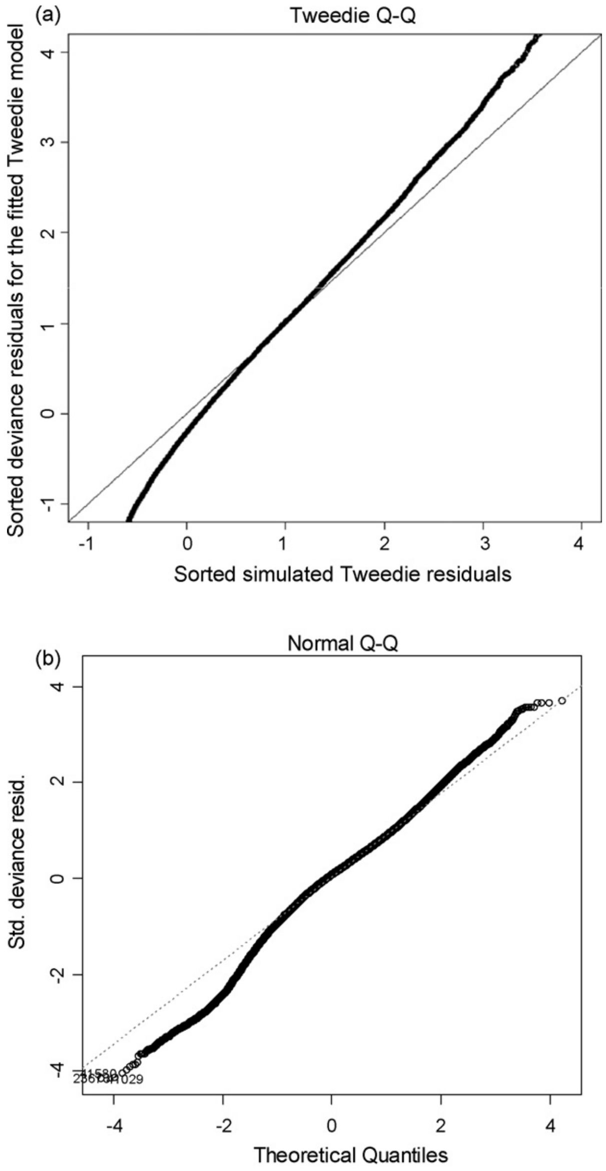

Fig. 3. Quantile–quantile (QQ) plots of the deviance residuals in the Tweedie model (a) and standard residuals in the ad hoc method (b) for yellowfin tuna.

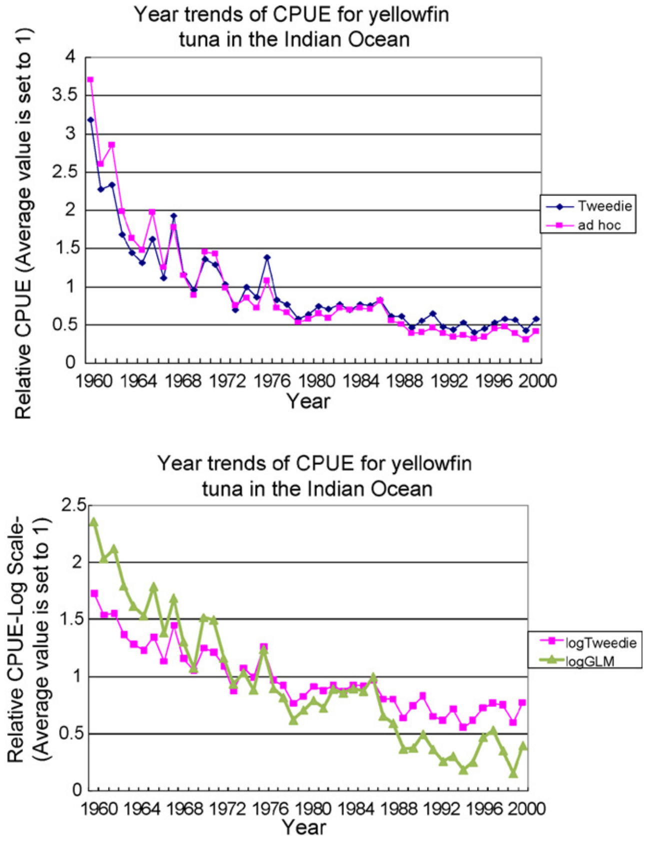

Fig. 4. Year trends of standardized CPUE obtained from the Tweedie distribution model and ad hoc method for yellowfin tuna in the Indian Ocean.

Table 2
Model comparison based on the results of 5-fold cross-validation for the example of yellowfin tuna

Candidate model	Pearson's correlation	Square error
Tweedie model	0.493920	3,749,210
Ad hoc method	0.468231	4,222,409

(i.e., sub-datasets) with concealing the observed values are shown in Tables 3 and 4. Both correlation and square error in the Tweedie model are better than those in the ad hoc method (Tables 3 and 4). The correlation plots between the observed and the predicted CPUE values for the whole are illustrated in Fig. 5. Judging from the figure, the Tweedie model is a little more balanced about the observed CPUE and corresponding predicted one than the ad hoc method because it is difficult for the ad hoc method to predict the large observed CPUE values moderately.

Table 3
Pearson's correlation coefficient in each sub-dataset by 5-fold cross-validation in the example of yellowfin tuna

Correlation	I	II	III	IV	V
Tweedie model	0.532	0.473	0.486	0.509	0.486
Ad hoc method	0.508	0.437	0.458	0.498	0.461

Table 4
Square error in each sub-dataset by 5-fold cross-validation for the example of yellowfin tuna

Square error	I	II	III	IV	V
Tweedie model	578,187	864,574	798,697	615,102	902,651
Ad hoc method	652,065	972,004	910,298	685,552	1,002,491

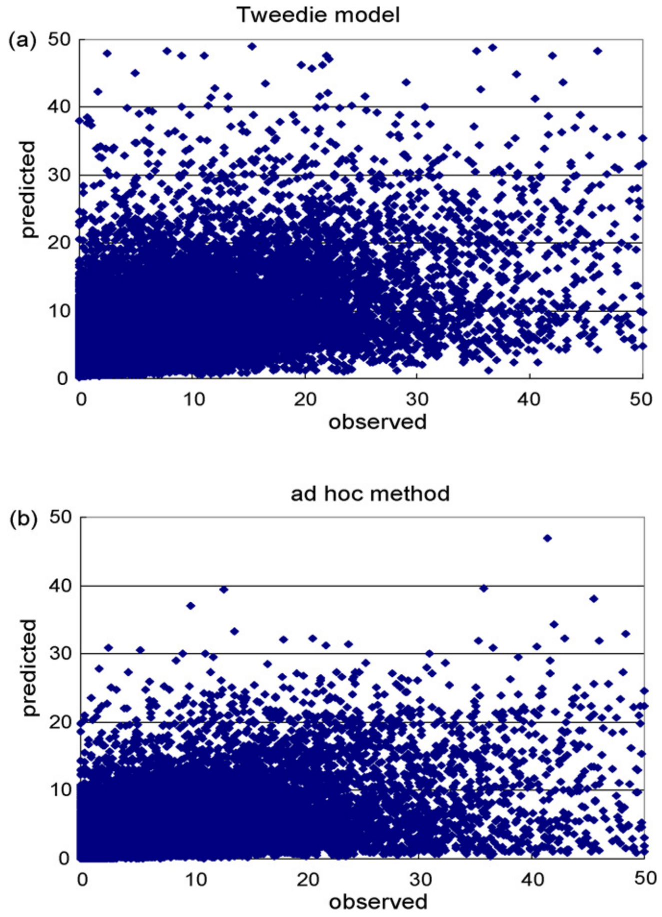

Fig. 5. Overall correlation plots of the observed and the predicted CPUE in the Tweedie model (a) and in the ad hoc method (b) for yellowfin tuna.

3.2. Case study 2: silky shark in the North Pacific Ocean by Japanese training longline vessels

As a result of the model selection of many candidate models in the range of null model to full model using BIC, same explanatory factors were selected in the ad hoc method and Catch-NB model (formulae (13) and (14)). Therefore, this combination of effects was also used in the Tweedie distribution and the Delta-lognormal model (both in the 1st and 2nd step).

Ad hoc method

$$\begin{aligned}\log(\text{CPUE}+0.01) &= \text{Intercept}+\text{Year}+\text{Area}+\text{Quarter}+\text{Gear}\\ &\quad +\text{Area}*\text{Gear}+\text{Error},\quad \text{Error}\sim N(0,\sigma^2)\end{aligned} \tag{13}$$

Catch-NB model

$$\begin{aligned}E[\text{Catch}] &= \text{Effort}\times\exp(\text{Intercept}+\text{Year}+\text{Area}+\text{Quarter}+\text{Gear}\\ &\quad +\text{Area}*\text{Gear},\\ &\quad \text{Catch}\sim \text{NB}(\alpha,\beta),\ (\text{Effort is assumed as the offset})\end{aligned} \tag{14}$$

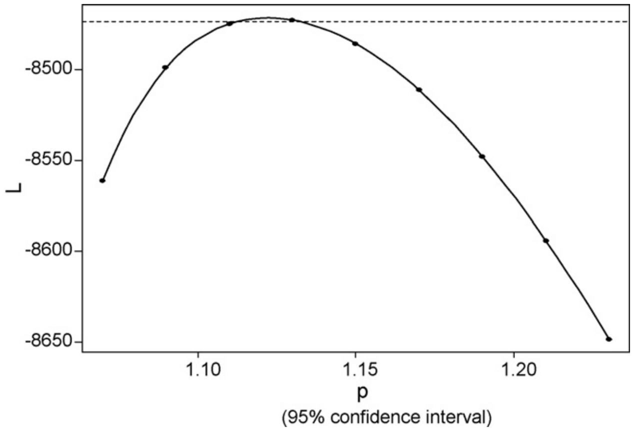

Fig. 6. Value of log-likelihood function (L) changing the power-parameter (p) in the Tweedie model for CPUE standardization of silky shark in the North Pacific Ocean caught by the Japanese longline training vessels.

Delta-lognormal model

1st step:

$$E\left[\log\left\{\frac{q}{1-q}\right\}\right] = \text{Intercept} + \text{Year} + \text{Area} + \text{Quarter} + \text{Gear} + \text{Area} * \text{Gear},$$

$$q(\text{ratio of zero-catch}) \sim \text{Binomial}(\theta) \quad (15)$$

2nd step:

$$\log(\text{CPUE}) = \text{Intercept} + \text{Year} + \text{Area} + \text{Quarter} + \text{Gear} + \text{Area} * \text{Gear} + \text{Error}, \quad \text{Error} \sim N(0, \sigma^2) \quad (16)$$

In the Delta-lognomal model, formulae (15) and (16) in the 1st step and 2nd step are applied to the whole data and positive data (i.e., Catch > 0), respectively.

Specific computational procedures in the case study, the method for n-fold cross-validation and indices (Pearson's correlation coefficient and square error) for model validation, are similar to those in the former case study 1 except for adding the Catch-NB distribution and Delta-lognormal model to the candidate models. We described the parameter estimation of the Tweedie distribution, correlation plots of the predicted and the observed values by 5-fold cross-validation, model comparison using Pearson's coefficient and square error.

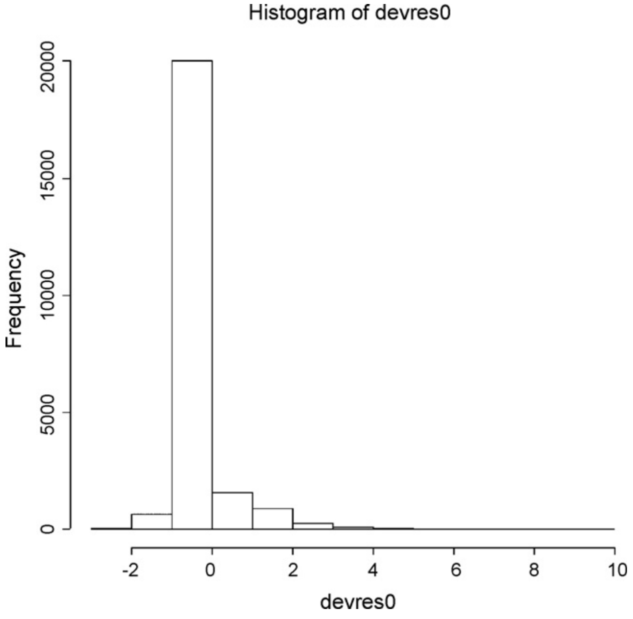

Fig. 7. Histogram of standard residual in the Tweedie model for silky shark.

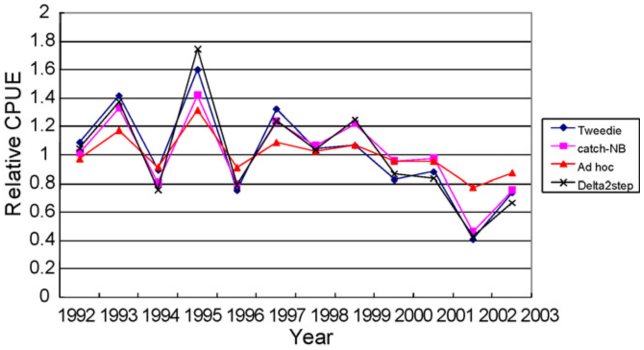

Fig. 8. Year trends of CPUE obtained from the Tweedie model, ad hoc method, delta-type two-step model and Catch-NB model for silky shark in the North Pacific Ocean.

Table 5
Dataset used for 5-fold cross-validation in the example of silky shark in the North Pacific Ocean caught by the Japanese longline training fishery

Sub-set	No. of data	Scenarios					
		Base case	I	II	III	IV	V
1	4688	Rule	C.V.	Rule	Rule	Rule	Rule
2	4687	Rule	Rule	C.V.	Rule	Rule	Rule
3	4688	Rule	Rule	Rule	C.V.	Rule	Rule
4	4687	Rule	Rule	Rule	Rule	C.V.	Rule
5	4688	Rule	Rule	Rule	Rule	Rule	C.V.

Rule and C.V. show the sub-dataset for rulemaking and cross-validation, respectively.

The values of log-likelihood function changing the power parameter (p) are shown in Fig. 6. The power parameter maximizing the log-likelihood was estimated about 1.12 and this imply the distribution pattern of the silky shark resembles the Poisson distribution. Fig. 7 shows the histogram of the standard deviance residuals and the fitting to the Normal distribution seems not to be so good. Fig. 8 shows CPUE year trends by LSMEANS based on Type III SS of the four candidate models, Tweedie model, ad hoc method, Delta-lognormal model and Catch-NB model. The year trends of standardized CPUE in ad hoc method are rather different from that of other three models.

Next, we describe the accuracy of the comparison of the Tweedie model, ad hoc method, Delta-lognormal model and Catch-NB model using 5-fold cross-validation. Table 5 shows the fifth sub-datasets divided randomly, in which "Rule" and "C.V." express the supervised data used for parameter estimation and unsupervised ones for cross-validation with concealing the observed values deliberately.

Tables 6–8 show the results of 5-fold cross-validation for the whole and in each sub-datasets, respectively. We carried out the

Table 6
Model comparison based on the results of 5-fold cross-validation for the example of silky shark

Candidate model	Pearson's correlation	Square error
Tweedie model	0.502957	6761.768
Catch-NB model	0.450111	11432.45
Delta-lognormal	0.484131	8065.108
Ad hoc method	0.446779	8814.842

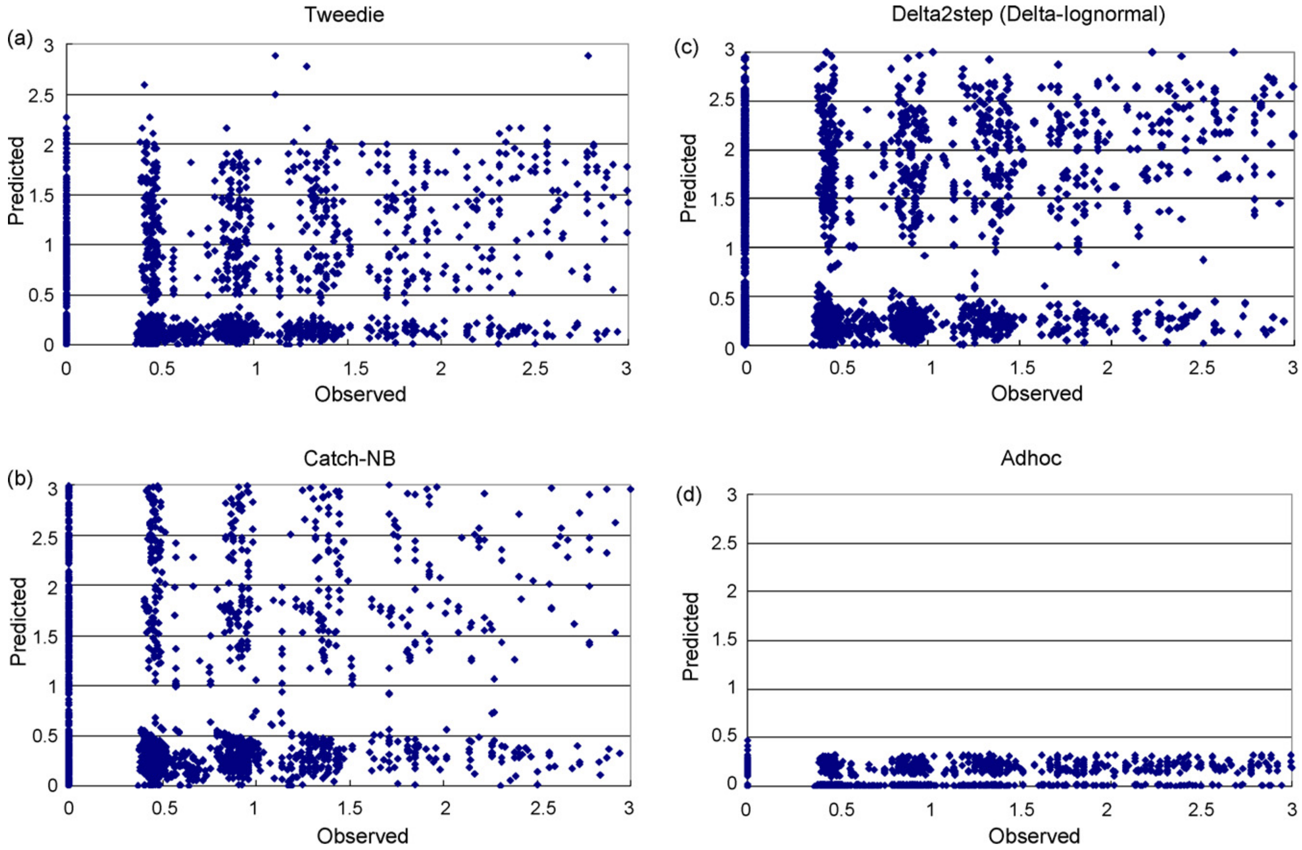

Fig. 9. Overall correlation plots of observed and predicted CPUE in the Tweedie model (a), in the Catch-NB model (b), in the Delta-lognormal model (c) and in the ad hoc method (d) for silky shark.

validation based on the Pearson's correlation coefficient and square error of the observed and the predicted values. In these tables, the Tweedie distribution model is the best and the Delta-lognormal model is the second best of the four candidate models judging from the value of both correlation coefficient and square error. In the other two models, the value of square error in the ad hoc method is smaller than that in the Catch-NB model. On the other hand, the value of square error in the Catch-NB model is wholly higher than that in the ad hoc method. Fig. 9 shows the correlation plots between the observed and the predicted values for the whole dataset in four models.

Table 7
Pearson's correlation coefficient in each sub-dataset by 5-fold cross-validation for the example of silky shark

Correlation	I	II	III	IV	V
Tweedie model	0.513	0.464	0.558	0.504	0.497
Catch-NB model	0.472	0.431	0.486	0.453	0.427
Delta-lognormal	0.482	0.462	0.515	0.485	0.49
Ad hoc method	0.456	0.422	0.47	0.46	0.443

Table 8
Square error each sub-dataset by 5-fold cross-validation for the example of silky shark

Square error	I	II	III	IV	V
Tweedie model	1277	1493	983	1166	1842
Catch-NB model	2035	2465	1911	2259	2762
Delta-lognormal	1527	1785	1268	1491	1994
Ad hoc method	1688	1846	1375	1517	2389

In the Tweedie model, predicted CPUE values seem to be a little small compared with the corresponding observed ones as a whole. The correlation plots are well-balanced rather than other three models. In Delta-lognormal model and Catch-NB model, the pattern of the correlation is rather similar and bias of the sign between the observed and the predicted CPUE is not so large. However, the difference of the absolute values between the observed and predicted CPUE is large especially in Catch-NB model. The ad hoc method has a bias that almost all of CPUE is estimated less than 0.5 regardless of the magnitude of the observed values. Values of square error in this method are small compared to the Catch-NB model and it seemed the cause that the most of the zero-catch data, which are approximately 85% of the total, is estimated to positive infinitesimal values. Therefore, model performance of ad hoc method seems to be inferior to that of Catch-NB model.

4. Discussion

As a result of the accuracy evaluation using *n*-fold cross-validation, where we used Pearson's correlation coefficient and square error between the observed and the predicted CPUE values, the model performance of the Tweedie model is better than that of other candidate models in both case studies.

In the former case study 1 in Sections 2.1 and 3.1, the ratio of zero-catch is about 10% and it is not high as the target tuna species. As a result of 5-fold cross-validation by correlation coefficient and square error, accuracy of the Tweedie model is slightly higher than that of ad hoc method and the difference is not so large. CPUE year trends of both models are rather similar. These are attributable to the lower rate of zero-catch. In other words, the superiority of the

Tweedie model does not appear so clearly if the rate of zero-catch is low. Judging from these analyses, it seems to be reasonable to apply the ad hoc method from the practical viewpoint if the ratio of zero-catch (R) is not so large (e.g., $R < 1/3$).

On the contrary, if the ratio of zero-catch is rather high in the latter case (described in Sections 2.2 and 3.2) which is a common ratio for by-catch species. As a result of n-fold cross-validation based on the indices of Pearson's correlation coefficient and square error, the performance of the Tweedie model is the best and that of the Delta-lognormal model is the second best. These models are rather superior to the ad hoc method and Catch-NB model.

In the comparison of the previously used methods, the ad hoc method and Catch-NB model, the Pearson's correlation coefficient and square error are better in the Catch-NB model and ad hoc method, respectively. However, the ad hoc method has a large bias because almost all of the estimated CPUE show less than 0.5 regardless of the magnitude of observed CPUE values.

The three candidate models, the Tweedie distribution, Delta-lognormal and Catch-NB model, show similar year trends of CPUE and the behavior of the ad hoc method is quite different. In fact, decreasing year trends were obtained from these three models to some degree or another although yearly CPUE based on the ad hoc model was rather stable.

Therefore, in the case that the ratio of zero-catch is high, we suggest to utilize the best Tweedie distribution model; otherwise the delta-type two-step method such as the Delta-lognormal model for practical reasons (i.e., the computation using the Tweedie model is generally more difficult than that through the Delta-lognormal model although R software is a powerful tool to analyze by the Tweedie distribution model) because the model performance is rather good compared to the ad hoc method and Catch-NB model. On the other hand, it is not so good to use the ad hoc method in such a case because this method has a large bias both for point estimation and for interval estimation.

Acknowledgements

The author thanks Dr. Hiroaki OKAMOTO and Mr. Hiroaki MATSUNAGA, National Research Institute of Far Seas Fisheries, for giving us processed catch and effort data utilized in the two case studies. Useful comments of two anonymous referees and Prof. Hiroe TSUBAKI, University of Tsukuba, are appreciated.

References

Anon., 1997. Report of the bluefin tuna methodology session. International Commission for the Conservation of Atlantic Tunas., Coll. Vol. Sci. Pap., vol. 40, 6(1), pp. 187–201.

Breiman, J.H., Friedman, R.A., Olshen, R.A., Stone, C.J, 1984. Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, p. 368.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information—Theoretic Approach. Springer, New York, p. 353.

Candy, S.G., 2004. Modeling catch and effort data using generalized linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. CCAMLR Science 11, 59–80.

Dobson, A.J., 1990. An Introduction to Generalized Linear Models. Chapman and Hall, London, p. 174.

Gavaris, S., 1980. Use of a multiplicative model to estimated catch rate and effort from commercial data. Canadian Journal of Fish and Aquatic Science 37, 2272–2275.

Jorgensen, B., 1997. The Theory of Dispersion Models. Chapman and Hall, London, p. 237.

Kawakita, M., Minami, M., Eguchi, S., Lennert-Cody, C.E., 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. Fisheries Research 76, 328–343.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.

Lo, N.C.L.D., Jacobson, L.D., Squire, J.L., 1992. Indices of relative abundance from fish spotter data based on Delta-Lognormal models. Canadian Journal of Fish and Aquatic Science 49, 2515–2526.

Maunder, M.M., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 74, 141–159.

Morris, C.M., 1982. Natural exponential families with quadratic variance function. Annals of Statistics 11, 59–67.

Reed, W.J., 1996. Analyzing catch-effort data allowing for randomness in the catching process. Canadian Journal of Fish and Aquatic Science 43, 174–186.

Robson, D.S., 1966. Estimation of the relative fishing power of individual ships. Research Bulletin, International Commission for the North-west Atlantic Fisheries 3, 5–14.

SAS, 2004. SAS/STAT 9.1 User's Guide, Volumes 1–7. SAS Institute Inc.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Smyth, G.K., 1996. Regression modeling of quantity data with exact zeroes. In: Proceedings of the Second Australia-Japan Workshop on Stohastic Models in Engineering Technology and Management, Technology Management Centre, University of Queensland, pp. 572–580.

Tweedie, M.C.K., 1984. An index which distinguishes between some important exponential families. In: Ghosh, J.K., Roy, J. (Eds.), Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Indian Statistical Institute, Calcutta, pp. 579–604.