CCAMLR Science, Vol. 11 (2004): 59–80

MODELLING CATCH AND EFFORT DATA USING GENERALISED LINEAR MODELS, THE TWEEDIE DISTRIBUTION, RANDOM VESSEL EFFECTS AND RANDOM STRATUM-BY-YEAR EFFECTS

S.G. Candy
Australian Antarctic Division
Channel Highway, Kingston 7050
Tasmania, Australia
Email – steve.candy@aad.gov.au

Abstract

The current standard method for modelling catch and effort data for Patagonian toothfish (*Dissostichus eleginoides*) for CCAMLR areas is to model the haul-by-haul ratios of catch to effort as the response variable in a generalised linear model (GLM) with a square-root link function and a unit variance function. A time series of standardised CPUE estimates and their precision can be obtained from the 'fishing year' parameter estimates together with 'baseline' parameter estimates, their variance–covariance matrix, and the inverse-link function. An alternative GLM with a more rigorous theoretical basis is introduced here. Catch is modelled as the response variable using a GLM with a power variance function, with the power parameter (λ) estimated using a profile extended quasi-likelihood, and a log link function with log of effort as an offset. For $1 < \lambda < 2$ this model is equivalent to assuming a compound Poisson-gamma distribution (i.e. Tweedie distribution) for catch that, unlike lognormal or gamma distributions, admits zero values.

In addition, random vessel effects are introduced into the GLM, as specified by a generalised linear mixed model (GLMM), in order to provide more efficient estimates of the standardised CPUE time series and more realistic estimates of their precision. Extra efficiency is gained by recovery of inter-vessel information as a result of the imbalance in the number of hauls in the year-by-vessel cross-classification. Further, the inclusion of an area stratum by fishing year interaction as an additional random effect in the GLMM is investigated. Fitting the stratum-by-year interaction as a fixed effect is problematic since it requires weighting of the individual stratum estimates by the areal extent of the stratum in order to obtain overall yearly standardised catch-per-unit-effort (CPUE) estimates. Without stratified random sampling, the determination of stratum areas that will give unbiased standardised CPUE estimates may be difficult. Fitting the stratum-by-year interaction as a random effect avoids this difficulty, and diagnostic methods to evaluate the validity of considering this interaction as random are described.

The methods are demonstrated using catch and effort data from two commercial *D. eleginoides* fisheries: the longline fishery around South Georgia (Subarea 48.3) and the trawl fishery around Heard Island and MacDonald Islands (Division 58.5.2).

Résumé

La méthode standard actuelle servant à modéliser les données de capture et d'effort de pêche de la légine australe (*Dissostichus eleginoides*) des secteurs de la CCAMLR consiste à modéliser les rapports trait par trait entre la capture et l'effort de pêche en tant que variable réponse dans un modèle linéaire généralisé (GLM) avec une fonction lien racine carrée et une fonction variance unité. Une série chronologique d'estimations normalisées de la CPUE et leur degré de précision peuvent être obtenus à partir des estimations paramétriques de «l'année de pêche» combinées à celles «de base», leur matrice de variance–covariance et la fonction de lien inverse. Un autre GLM est présenté ici, fondé sur une base théorique plus rigoureuse. La capture est modélisée en variable réponse par un GLM avec une fonction de variance de la puissance, dont le paramètre puissance (λ) est estimé par une quasi-probabilité au profil étendu et une fonction de lien log avec le log compensateur de l'effort de pêche. Pour $1 < \lambda < 2$, ce modèle équivaut à présumer une distribution composite Poisson-gamma (c-à-d. une distribution de Tweedie) de la capture qui, contrairement aux distributions lognormales ou gamma, admet les valeurs nulles.

De plus, les effets aléatoires des navires sont introduits dans le GLM, tells qu'ils sont spécifiés par le modèle linéaire généralisé mixte (GLMM), pour arriver à des estimations plus efficaces de la série chronologique de la CPUE normalisée et des estimations plus réalistes de leur précision. On gagne en efficacité par le recouvrement d'informations

entre les navires en raison du déséquilibre du nombre de poses dans la classification croisée année par navire. De plus, un effet aléatoire supplémentaire dans le GLMM est étudié : l'inclusion d'une interaction de la strate et de l'année de pêche. L'ajustement de cette interaction en effet fixe est problématique car cela requiert la pondération de l'estimation particulière de la strate par sa superficie pour obtenir des estimations générales annuelles normalisées de la capture par unité d'effort de pêche (CPUE). Sans l'échantillonnage aléatoire stratifié, il pourrait être difficile de déterminer quelles zones de strate donneraient des estimations normalisées non biaisées de la CPUE. Cette difficulté est levée par l'ajustement de l'interaction strate-année en tant qu'effet aléatoire; les méthodes diagnostiques permettant d'évaluer la validité de considérer comme aléatoire cette interaction sont décrites.

Les méthodes sont démontrées grâce aux données de capture et d'effort de pêche de deux pêcheries commerciales de *D. eleginoides* : la pêcherie à la palangre autour de la Géorgie du Sud (sous-zone 48.3) et la pêcherie au chalut autour des îles Heard et MacDonald (division 58.5.2).

Резюме

Принятый стандартный метод моделирования данных по уловам и усилию для патагонского клыкача (*Dissostichus eleginoides*) в районах АНТКОМа заключается в моделировании соотношения уловов и усилия по каждой отдельной выборке как зависимой переменной в обобщенной линейной модели (GLM) с функцией связи в виде квадратного корня и функцией единичной дисперсии. Временной ряд стандартизованных оценок CPUE и их точность можно получить по оценкам параметра «промысловый год», а также оценкам «базовых» параметров, их ковариационной матрице и функции обратной связи. Представлена альтернативная GLM с более строгим теоретическим основанием. Улов моделируется как зависимая переменная на основе GLM со степенной функцией дисперсии, порядок степени которой (λ) рассчитывается на основе квази-правдоподобия с расширенным профилем, и логарифмической функции связи с коррекцией на логарифм усилия. Для $1 < \lambda < 2$ данная модель эквивалентна предположению о сложном Пуассон-гамма распределении (т.е. распределении Твиди) улова, которое, в отличие от логнормального или гамма-распределения, допускает нулевые величины.

Кроме того, в GLM введены случайные эффекты для судов, определенные по обобщенной линейной смешанной модели (GLMM), с целью получения более эффективных оценок стандартизованных временных рядов CPUE и более реалистичных оценок их точности. Дополнительная эффективность достигается путем восстановления информации по судам в результате несбалансированности количества выборок в перекрестной классификации годов по отдельным судам. Кроме того, исследуется включение в GLMM взаимодействия между районами и годами промысла в качестве дополнительного случайного эффекта. Подбор взаимодействия зон по годам как фиксированного эффекта является проблематичным, т.к. для получения суммарной годовой стандартизованной оценки улова на единицу усилия (CPUE) необходимо взвешивание оценок по отдельным зонам на площадь этих зон. Без стратифицированной случайной выборки трудно определить площадь зон для получения несмещенных оценок стандартизованных CPUE. Включение взаимодействия зона-годы как случайного эффекта позволяет избежать этой трудности; описываются диагностические методы для оценки обоснованности рассмотрения этого взаимодействия как случайного.

Для иллюстрации этих методов используются данные по уловам и усилию, полученные в ходе двух коммерческих промыслов *D. eleginoides*: ярусного промысла в районе Южной Георгии (Подрайон 48.3) и тралового промысла в районе о-вов Херд и Макдональд (Участок 58.5.2).

Resumen

El método estándar utilizado actualmente para modelar los datos de captura y esfuerzo de la pesca de austromerluza negra (*Dissostichus eleginoides*) en áreas de la CCRVMA consiste en representar el cociente entre la captura de un lance y el esfuerzo correspondiente, como la variable de respuesta en un modelo lineal generalizado (GLM), con una función de enlace de raíz cuadrada y una función de varianza unitaria. Se puede obtener una serie cronológica de valores del CPUE normalizado y de su precisión a partir de un conjunto

de estimaciones del parámetro "año de pesca" y de parámetros "básicos", su matriz varianza-covarianza y el inverso de la función de enlace. Aquí se presenta otro modelo GLM apoyado en una teoría más rigurosa. Se utiliza la captura como variable de respuesta en un modelo GLM con una función de potencia para la varianza, estimando la potencia (λ) con un estimador de cuasi-verosimilitud para ampliar el perfil, y una función de enlace logarítmica compensada por el logaritmo del esfuerzo. Para $1 < \lambda < 2$, el modelo equivale a suponer que la captura tiene una distribución Poisson-gamma (i.e. distribución Tweedie) que, a diferencia de las distribuciones lognormal o gamma, admite valores cero.

También se introducen al GLM los efectos aleatorios del factor barco, de acuerdo con un modelo lineal mixto generalizado (GLMM), para obtener mejores estimaciones de la serie cronológica del CPUE normalizado y estimaciones más realistas de su exactitud. Se aumenta la eficacia del modelo al recuperar información de distintos barcos, como resultado de la variabilidad en el número de lances al clasificar éstos por año y barco. Se estudia además la inclusión de una interacción entre estratos y año de pesca, como variable aleatoria adicional en el GLMM. Es difícil ajustar la interacción estrato-año como efecto fijo, ya que esto requiere una ponderación de las estimaciones de los estratos individuales por la extensión geográfica del estrato para calcular los totales anuales de captura por unidad de esfuerzo (CPUE) normalizados. Sin un muestreo aleatorio estratificado, es difícil determinar las áreas de estrato que darían estimaciones sin sesgos del CPUE. Al incorporar la interacción estrato-año como variable aleatoria se elimina esta dificultad. Se describen los métodos de diagnóstico para evaluar si es válido considerar que esta interacción ocurre al azar.

Se utilizan los datos de captura y esfuerzo de dos pesquerías comerciales dirigidas a *D. eleginoides*: la pesquería de palangre alrededor de Georgia del Sur (Subárea 48.3) y la de arrastre alrededor de las islas Heard y MacDonald (División 58.5.2), para demostrar estos métodos.

Keywords: *Dissostichus eleginoides*, catch-per-unit-effort, generalised linear mixed models, Tweedie distribution, South Georgia, Subarea 48.3, Heard and MacDonald Islands, Division 58.5.2, CCAMLR

Introduction

Since 1995, standardisation of commercial catch-per-unit-effort (CPUE) data for the Patagonian toothfish (*Dissostichus eleginoides*) fishery in CCAMLR Subarea 48.3 has followed a standard methodology (SC-CAMLR, 2002). This method aims to provide, as an index of abundance, a time series of standardised CPUE estimates, where standardisation is carried out using predictions from a fitted generalised linear model (GLM) (McCullagh and Nelder, 1989) with haul-level CPUE values as the response variable. CPUE is calculated as the ratio of the catch (kg) and the number of hooks deployed for a haul. The linear predictor of the GLM incorporates, in addition to YEAR (a categorical factor representing the years of fishing), categorical and continuous predictors which influence CPUE and may include their interactions. The partial regression coefficients for YEAR, when added to 'baseline' values defined using parameter estimates for reference levels of other predictors, after being suitably transformed by the inverse-link function, provide the standardised CPUE series.

This paper discusses the current and alternative approaches to modelling catch and effort data in order to provide a series of standardised CPUE estimates along with their confidence intervals. This includes the choice of response variable (i.e. either CPUE or catch), the link function and variance function for GLMs, and the incorporation of random effects for vessels and YEAR interactions in the linear predictor of the GLM to give a generalised linear mixed model (GLMM) (Schall, 1991; Breslow and Clayton, 1993; Diggle et al., 1994). Interactions between YEAR and other factors must be averaged out of the linear predictor in order for the YEAR regression coefficients to be interpretable as overall indices of relative abundance for the fishery. This averaging operation is straightforward if these interactions are fitted as random effects.

The statistical methods are demonstrated using longline catch and effort data for CCAMLR Subarea 48.3 and trawl catch and effort data for CCAMLR Division 58.2.2 up to and including 2002 and 2003 fishing years respectively.

Methods

GLMs and GLMMs for CPUE and catch

The CCAMLR standard method fits the following GLM

$$y_{ij} = \frac{C_{ij}}{E_{ij}} = \eta_{ij}^2 + \varepsilon_{ij}, i = 1,...,q; \quad j = 1,...,n_i \qquad \text{model (1)}$$

where y_{ij} is the observed value of the response variable (Y) calculated as the ratio of the catch (C) to effort E for the jth haul from the ith vessel, η is the linear predictor, and ε_{ij} is an independently and identically distributed random error with dispersion parameter (ϕ) corresponding to σ^2 in the usual linear model (LM) notation. The catch (C) is the total haul weight of fish (kg) and E_{ij} is defined as either the number of hooks on the jth deployment of a longline or the area trawled in the jth haul from the ith vessel. The linear predictor is the sum

$$\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_p x_{pij}$$

or in matrix terms is

$$\eta = \mathbf{X}\beta$$

where the x's are predictor variables and the β's are the corresponding regression coefficients or parameters.

For a GLM the expected value of the response is related to the linear predictor via the link function, $g(.)$ so that

$$g(\mu) = \eta$$

where $E(Y|\eta) = \mu$ is the expected value of the response and the variance is given by

$$Var(Y|\eta) = \phi V(\mu)$$

where $V(.)$ is a variance function involving known parameters and ϕ is a dispersion parameter.

In model (1) the response variable is catch rate or CPUE and the link function is the square-root function. The distribution of catch rate can be difficult to model since it is the ratio of two random variables. Another approach is to condition on the observed values of effort and model haul-by-haul catch as the response variable. Using a log-link function (i.e. $g(.) = \log_e(.)$) an alternative GLM for catch rate can therefore be expressed as

$$y_{ij} = C_{ij} = \exp\left[\log_e\left(E_{ij}\right) + \eta'_{ij}\right] + \varepsilon'_{ij}, i = 1,...,q; j = 1,...,n_i \qquad \text{model (2)}$$

where $\log_e(E_{ij})$ is called an 'offset' and where the linear predictor is then the sum of two components, $\eta_{ij} = \log_e\left(E_{ij}\right) + \eta'_{ij}$. A convenient class of variance functions for modelling catch data is the power variance function $V(\mu) = \mu^\lambda$ with power parameter λ in the range $1 \geq \lambda \geq 2$. The power parameter must be assumed to be fixed and known for the purposes of the standard GLM formulation, but a method of estimating λ using a profile extended quasi-deviance (PEQD) (McCullagh and Nelder, 1989) will be given. The lower and upper extremes of the above range for λ correspond in terms of quasi-likelihood to Poisson and gamma distributions respectively, while values of λ within the range correspond to a compound Poisson distribution (i.e. Tweedie distribution) for catch.

In model (1) the link function is the square-root function and this model can be expressed with catch as the response variable so that

$$y_{ij} = C_{ij} = \left(E_{ij}^{0.5}\eta'_{ij}\right)^2 + \varepsilon'_{ij}, i = 1,...,q; \quad j = 1,...,n_i \qquad \text{model (3)}$$

where $\varepsilon'_{ij} = E_{ij}\varepsilon_{ij}$ and $Var(\varepsilon'_{ij}) = \phi_{ij}$ with corresponding dispersion model $\phi_{ij} = \phi E_{ij}^2$.

For model (3) the link function is still the square root but all terms in the linear predictor are multiplied by $E_{ij}^{0.5}$; however, the variance function cannot be expressed in the usual GLM form which must be a function only of μ (i.e. including the simplest case of $V(\mu) = \mu^0 = 1$ corresponding to a Gaussian error distribution) and known constants (i.e. variance parameters). Model (3) can be fitted as a double generalised linear model (DGLM) (Smyth and Verbyla, 1999) as described in Candy (2003a). However, in order to allow direct comparison of model (2) with model (3), which both use catch as the response variable, it is necessary to specify the same variance function for both. Therefore, the variance of catch in model (3) was also assumed to be $Var(Y|\eta) = \phi\mu^\lambda$.

Further, by noting that the log-link function can be imbedded in the power family of link functions where the power parameter is θ, both models (2) and (3) can be generalised to give

$$y_{ij} = C_{ij} = \left(E_{ij}^{\theta}\eta'_{ij}\right)^{\frac{1}{\theta}} + \varepsilon'_{ij}, i = 1,...,q;$$
$$j = 1,...,n_i \qquad \text{model (4)}$$

The log-link function falls within this family of power-link functions when $\theta = 0$ while the square-root link function corresponds to $\theta = 0.5$.

Fitting a GLM with Tweedie distributed errors using quasi-likelihood

The Tweedie distribution (Jørgensen, 1997) is the exponential family dispersion model with variance function given by the power function with $1 < \lambda < 2$. From quasi-likelihood theory the quasi-deviance, $D(y,\mu)$, where

$$D(y,\mu) = -2\sum_{i,j}\int_{y_{ij}}^{\mu_{ij}} \frac{(y-t)}{V(t)}dt,$$

corresponding to the power variance function with $\lambda = 0, 1, 2$ gives the deviance for normal (i.e. Gaussian), Poisson, and gamma distributions respectively. The maximum quasi-likelihood estimate of β is that which minimises $D(y,\mu)$ and is obtained using the iteratively weighted least-squares GLM fitting algorithm (McCullagh and Nelder, 1989).

When $1 < \lambda < 2$ the distribution of $Y = y$ is intermediate between a Poisson and gamma distribution whereby $Z = Y$ and Z is defined as

$$Z = W_1 + W_2 + ...W_k + ... + W_N$$

where the W_k (k=1,…,N) are independently and identically distributed gamma-random variables with mean μ_w and variance $\phi_w\mu_w^2$ (e.g. the weight of the kth fish in the haul) and N is distributed as a Poisson random-count variable with mean τ (e.g. the number of fish in the haul). Since N can be zero at a frequency determined by the Poisson distribution with mean τ, the distribution of Z has non-zero mass at $Z = 0$. The distribution of Z is called a compound Poisson or Poisson-gamma (CPG) distribution and has also been called a Tweedie distribution (Jørgensen, 1997) (Appendix 1 gives its probability density function). From quasi-likelihood theory (Appendix 1), fitting the Tweedie distribution to catch, conditional on effort and predictor variables specified in η'_{ij}, can be carried out by fitting a GLM (e.g. with systematic component specified by model (2) or (4)) with the power variance function. This requires estimation of ϕ and λ in addition to β, where λ is estimated by profiling an extended definition of quasi-likelihood as described below. First, using a similar profile estimation method, estimation of the link power parameter (θ) is described.

Estimation of power parameters and the dispersion parameter

Given a starting value of λ that models the conditional variance $Var(Y|\eta)$ reasonably well, an optimal value for θ in terms of fit to the catch data can be obtained using a profile (quasi-) deviance (PQD), D, given for model (4) by

$$D\left(\theta\middle|y_{ij},E_{ij},\hat{\eta}'_{ij},\lambda=\lambda_0\right) =$$
$$D\left(\theta\middle|y_{ij},\hat{\mu},\lambda=\lambda_0\right) = \sum_{i,j} d_{ij} \qquad (5)$$

where the formula for the deviance contribution, d_{ij}, for the power variance function for $1 < \lambda < 2$ is given in Appendix 2. Since the value of λ is fixed for all values of θ in equation (5), the PQD can be obtained for a grid of values for θ. The optimal value of θ is that for which the PQD is a minimum.

Given an estimate of θ, an empirical estimate of λ can be obtained using a profile extended (quasi-) deviance (PEQD), D^*, for a given grid of values for λ where D^* is given by equation (10.3) of McCullagh and Nelder (1989) as

$$D^*\left(\lambda\middle|y_{ij},E_{ij},\hat{\eta}'_{ij},\theta=\theta_0\right) =$$
$$\sum_{i,j}\frac{d_{ij}}{\phi_{ij}} + \sum_{i,j}\log_e\left\{2\pi\phi_{ij}V\left(y_{ij}\right)\right\} \;;\; y_{ij} > 0 \qquad (6)$$

where for the specific case of model (4) $\phi_{ij} = \phi$ and $V(y_{ij}) = y_{ij}^{\lambda}$. It should be noted that although fitted values of y_{ij} given by $\hat{\mu}_{ij}$ are always greater than zero, the zero values of y_{ij} must be excluded from the second term in equation (6) since $V(y_{ij}) = 0$ for these zero values (however, see below for the contribution of these zero values through ϕ). In order to calculate equation (6), an estimate of ϕ is required. The estimate can either be the Pearson chi-square statistic divided by the residual degrees of freedom or the residual mean deviance (McCullagh and Nelder, 1989). If the latter is used then

$$\hat{\phi} = \frac{1}{N-p}\sum_{i,j} d_{ij}$$

where $N = \sum_i n_i$, the total number of observations, and p is the dimension of $\hat{\beta}$. Therefore equation (6) simplifies to

$$D^*\left(\lambda\middle|y_{ij},E_{ij},\hat{\eta}'_{ij},\theta=\theta_0\right) =$$
$$(N-p) + \sum_{i,j}\log_e\left\{2\pi\hat{\phi}\,V\left(y_{ij}\right)\right\} \;;\; y_{ij} > 0\,. \qquad (7)$$

Note that zero values of y_{ij} contribute to equation (7) through $\hat{\phi}$ via their deviance contributions.

Random effects

Random vessel effects model the variation in catch, conditional on effort and the other fixed terms in the linear predictor, due to variation between vessels in their ability to catch fish which will depend on the attributes of the vessel, its crew, and the total extent of fishing grounds that they target. For the remainder it will be assumed that catch weight is Tweedie-distributed conditional on any random effects in the linear predictor.

In mixed-model terminology the x's in the linear predictor introduced earlier are called fixed effects to distinguish them from random effects. The GLMM formulation that incorporates random effects in the linear predictor gives a conditional linear predictor

$$\eta_u = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

where $\mathbf{u}$ is a vector of random effects distributed as multivariate normal with expected value zero and variance–covariance matrix Σ, and $\mathbf{Z}$ is the random-effect design matrix. The specification of the random effects term $\mathbf{Zu}$ is general enough to allow for more than one sampling level (multilevel models) and more than one set of correlated random effects per sampling level (Candy, 2000).

The particular GLMM employed here (i.e. only the log link was considered) with a single random effect term which generalises model (2) is given by

$$\begin{aligned} y_{ij} = C_{ij} &= \exp\left[\log_e\left(E_{ij}\right) + \eta'_{ij} + u_i\right] \\ &+ \varepsilon''_{ij}, i = 1,\ldots,q; j = 1,\ldots,n_i \end{aligned} \tag{8}$$

where u_i is a Gaussian random effect (e.g. vessel effect) where $E(u_i) = 0, Var(u_i) = \sigma_u^2$ and $Cov(u_i u_{r\neq i}) = 0$. The conditional error, ε'', has variance given by

$$Var\left(\varepsilon''_{ij}\right) = \phi\mu^{\lambda}_{(u)ij}$$

where $\mu_{(u)} = g^{-1}\left(\eta_{(u)}\right)$ and $\eta_{(u)ij} = \log_e\left(E_{ij}\right) + \eta'_{ij} + u_i$.

This is the simplest GLMM possible as it incorporates only a random intercept. Equation (8) can be fitted using penalised quasi-likelihood (PQL) (Schall, 1991; Breslow and Clayton, 1993). Once equation (8) has been fitted, 'population-average' (PA) estimates (Zeger et al., 1988) of the β can be obtained by integrating u out of the conditional predictions of catch (i.e. 'averaging out' $\mathbf{u}$). Using analytical integration this gives

$$E\left(Y \mid \eta', E\right) = \exp\left[\log_e(E) + \eta' + \tfrac{1}{2}\sigma_u^2\right].$$

Since the model is log-linear any chosen reference level for an additive factor included in η' other than YEAR simply scales the estimated standardised CPUE series by a constant fraction. The same is true for the scaling factor $\exp\left(\frac{1}{2}\sigma_u^2\right)$. Averaging out other random effects such as cruise-within-vessel or stratum-by-year interaction, where stratum may be a subdivision of the ocean area of the fishery (Brandão et al., 2002) or a depth stratum or a combination of both (Punt et al., 2000), can be handled in the same way. A uniform scaling of the standardised CPUE series does not affect its interpretation as a relative measure of abundance. Note that the estimates that are unscaled by $\exp\left(\frac{1}{2}\sigma_u^2\right)$ correspond to those obtained by setting u to its expected value, rather than integrating it out of the conditional expectation function, to give the so-called 'subject-specific' (SS) parameter estimates (Zeger et al., 1988). This simple scale difference between SS and PA estimates for log-link GLMMs only holds for the simple random intercept model (Grömping, 1996). Approximate best linear unbiased predictions (BLUPs) (Robinson, 1991) of the random effects, given by $\hat{\mathbf{u}}$, can be obtained from the fit of the GLMM.

The extended quasi-deviance (see equation (3.1) of Lee and Nelder, 2001) for equation (8) that extends the PEQD (7) to GLMMs is given by

$$\begin{aligned} D_u^*\left(\lambda \middle| y_{ij}, E_{ij}, \hat{\eta}'_{ij}, \theta = \theta_0\right) &= \sum_{i,j} \frac{d_{ij}\left(\hat{\mu}_{(u)}\right)}{\phi} \\ &+ \sum_{i,j} \log_e\left\{2\pi\phi V\left(y_{ij}\right)\right\} + q\log_e\left\{2\pi\sigma_u^2\right\} + r \end{aligned}$$

where $d_{ij}\left(\hat{\mu}_{(u)}\right)$ is the conditional deviance, $\hat{\mu}_{(u)} = \exp\left[\log_e\left(E_{ij}\right) + \hat{\eta}'_{ij} + \hat{u}_i\right]$, and $r = q - \frac{1}{\hat{\sigma}_u^2}\sum_i \hat{u}_i^2$ is the 'effective degrees of freedom'. It can be seen from r that, unlike fixed-effect parameters, the estimation of $\hat{\mathbf{u}}$ in the GLMM results in the loss of less than q degrees of freedom. In addition, $\hat{\mathbf{u}}$ is estimated in a separate step of the GLMM fitting algorithm to the generalised least-squares estimation step used to obtain $\hat{\beta}$ making the simultaneous estimation of $\hat{\mathbf{u}}$ and $\hat{\beta}$ feasible and stable even when q is large. This is often not the case when the u's are fitted as fixed effects.

Estimation of the dispersion parameter ϕ as the residual mean deviance

$$\hat{\phi} = \frac{1}{N - p - r}\sum_{i,j} d_{ij}\left(\hat{\mu}_{(u)}\right)$$

gives

$$D_u^*\left(\lambda\middle|y_{ij},E_{ij},\hat{\eta}'_{ij},\theta=\theta_0\right)=N-p + \sum_{i,j}\log_e\left\{2\pi\hat{\phi}V\left(y_{ij}\right)\right\}+q\log_e\left\{2\pi\hat{\sigma}_u^2\right\}. \quad (9)$$

The estimate of σ_u^2 is obtained using residual maximum likelihood (REML) (Schall, 1991; Breslow and Clayton, 1993).

Finally, approximate confidence intervals for the standardised CPUE series are obtained using an estimate of the variance of the sum of linear predictor terms used to form the CPUE estimates calculated using the variance–covariance matrix of the GLM or GLMM (fixed effect) parameter estimates, calculating *t*-statistics using the CPUE estimates on the linear predictor scale, and back-transforming the confidence interval end-points using the inverse link function.

Catch and effort data

Longline fishery for *D. eleginoides* in Subarea 48.3

The longline catch and effort data used for model fitting and testing came from 15 712 hauls from 54 vessels (SHIP), and 15 fishing years (1987 to 2002 excluding 1990) and included 294 zero values of catch (Candy, 2003a).

The standard CCAMLR method for modelling catch and effort for the longline fishery in Subarea 48.3 employs two predictive categorical factors besides YEAR: the nationality of the fishing vessel and the depth of set of the longline classified into one of the four categories: 0–500 m, 500–1 000 m, 1 000–1 500 m and 1 500 m and above. Other predictor variables, such as bait type, month etc., have been explored but found to not significantly improve the fit. The random effect term included in the GLMM was SHIP.

Trawl fishery for *D. eleginoides* in Division 58.5.2

The data used for model fitting and testing came from 3 354 hauls from 3 vessels over 37 cruises, distributed across 15 area strata and 7 fishing years (1997 to 2003) (Candy, 2003b). The data were as usual highly unbalanced and included 74 zero values of catch. Effort was measured as swept area (ha) (i.e. 'NetWingWidth' x 'TowDistance'/10 000) and only hauls which had a reliability score for TowDistance of 1 or 2 were used in the analyses. The random effects included in the GLMM were cruise and stratum-by-year interaction. These two random-effect terms were assumed to be independent. A cruise can only occur in a single year but may fish a number of strata, making it difficult to verify the independence assumption.

Note that all of the following models of haul-by-haul catch or CPUE have been fitted to the combined data of zero and non-zero catches.

Results

Longline fishery for *D. eleginoides* in Subarea 48.3

The fit of the standard model (1) in terms of ordinary residuals (i.e. observed minus fitted values) versus fitted values and sorted residuals versus normal quantiles is shown in Figure 1.

Fitting the log-link GLMM (8) with the profile extended quasi-deviance (PEQD) given by equation (9) is shown in Figure 2 for a grid of λ values. The corresponding graph (not shown) obtained using equation (6) for the GLM (2) was very similar (Candy, 2003a). The minimum extended quasi-deviance (EQD) occurs at approximately $\lambda = 1.3$ for both the GLM and GLMM.

Figure 3 shows the profile quasi-deviance (PQD) (5) for $\lambda_0 = 1.3$ for a grid of values of θ. The minimum quasi-deviance (QD) estimate of θ from Figure 3 is –0.4. The scaled PQD given by $D/\hat{\phi}$ is approximated by a χ_1^2 allowing a 95% support interval about $\hat{\theta}$ to be constructed. This interval is shown in Figure 3 as (–0.55, –0.2). It is clear from Figure 3 that the square-root link function (i.e. $\theta = 0.5$) gives a much poorer fit to the catch data than either the log or optimum power link functions (i.e. $\theta = 0$, –0.4 respectively). However, it was found that the optimum power-link function, θ = –0.4, gave unreasonably large standardised CPUE estimates for 1988 and 1993 (Candy, 2003a). Therefore, only the log-link function was pursued as an alternative to the square root.

Initially the GLMM (8) was fitted to the data from 57 vessels, but the random-effects estimates for two vessels were judged to be outliers and all hauls for these two vessels and the single haul for a third vessel were removed from the dataset to give 15 712 hauls from 54 vessels (Candy, 2003a). Figure 4 shows the estimated random SHIP effects ($\hat{u}_i$) for the reduced dataset as sorted values versus normal quantiles and as indexed values with approximate standard error bars. The estimate of the random-effect variance ($\hat{\sigma}_u^2$) was 0.06368 with a standard error of 0.01477. The estimate of $\hat{\phi}$ was 68.423 (s.e. 0.774).

Figure 5 shows a plot of conditional Pearson residuals (i.e. $\left(y-\hat{\mu}_{(u)}\right)\Big/V^{0.5}\left(\hat{\mu}_{(u)}\right)$) and conditional deviance residuals (i.e. $sign\left(y-\hat{\mu}_{(u)}\right)d_{ij}^{0.5}\left(\hat{\mu}_{(u)}\right)$) (McCullagh and Nelder, 1989) from the fit of GLMM (8) (λ = 1.3) versus conditional fitted values ($\hat{\mu}_{(u)}$) and sorted residuals versus normal quantiles.

Figure 6 displays the ordinary (conditional) residuals compared to a set of simulated Tweedie residuals. The set of simulated Tweedie (ordinary) residuals was obtained by generating a single simulated catch for each fitted value from equation (8) and then subtracting the fitted values from the simulated set of catches. Figure 7 shows the results of an extension to the method used to construct Figure 6. Instead of a single simulation set of Tweedie residuals, 100 random sets were generated and after sorting each set in the order of the corresponding observed residuals and then sorting across the 100 sets for each fitted value, the top and bottom two values of the 100 simulations were used to define a 96% confidence interval about the mean of the 100 simulations. Joining these confidence interval endpoints across the corresponding 15 712 observed residuals gives a 96% confidence region. This procedure was carried out for GLM (2) and GLMM (8) with Tweedie residuals and also for the square-root link model (1) and simulated normal residuals with variance given by $\hat{\phi}$ = 0.04857.

The marginal predictions from GLMM (8) (i.e. setting $\mathbf{u}_{vessel} \equiv 0$) were used to obtain a standardised CPUE series with reference levels of nationality = 'CHL', and depth.class = '1000_1500'. Figure 8 compares the standardised CPUE series estimates and their 95% confidence intervals that were obtained from each of the three models, the glm(link = sqrt) (model(1)), glm(link = log) (model (2), (λ = 1.3)) and glmm(link = log) (model (8), (λ = 1.3)).

Trawl fishery for *D. eleginoides* in Division 58.5.2

The marginal predictions from the GLMM (i.e. setting $\mathbf{u}_{cruise} \equiv 0$ and $\mathbf{u}_{SxY} \equiv 0$) were used to obtain a standardised CPUE series with reference levels obtained for fixed effects of Net.Type = 'Albatross' and Month = 'Apr'. The series for each of three strata is given in Figure 9 (Candy, 2003b). To see the effect of the random stratum-by-year (*S*x*Y*) interaction on the series, the estimated random effects for this term were added to the linear predictor to give conditional CPUE estimates. These are shown with corresponding approximate 95% confidence bars in Figure 9. Where the particular stratum-by-year combination did not occur in the data these confidence bars are missing in Figure 9 with the corresponding random effects set to zero for the purposes of this figure. For other results relating to the estimated random effects, the estimates for missing combinations of stratum and year were not set to zero but were instead excluded from calculations to avoid 'polluting' results by replacing unestimable values with zeros. Stratum was included as a fixed effect in the GLMM which explains why the standardised marginal CPUE estimates vary by stratum in Figure 9.

The estimates of the random-effect variances were 0.1119 and 0.2599 for stratum-by-year and cruise respectively. The estimate of the dispersion parameter ($\hat{\phi}$) was 30.32 for $\hat{\lambda} = 1.6$. For comparison with the results for the longline data from Subarea 48.3, setting λ = 1.3 gave a corresponding estimate of $\hat{\phi}$ = 357.07 compared to the much smaller estimate of 68.42 for the Subarea 48.3 data. Clearly either the trawl catch data is far more variable than the longline data, the effort as measured by swept area is a much poorer determinant of catch than hooks deployed, or both sources of variability are greater for the trawl data.

Figure 10 shows normal quantile–quantile plots for estimated random effects from the fit of the GLMM for: (a) random cruise effect estimates, and (b) random stratum-by-year (*S*x*Y*) effect estimates. Figure 11 shows a histogram of the *S*x*Y* random-effect estimates. Figure 12 shows random stratum-by-year random-effect estimates (*S*x*Y*_re) where estimates for the same stratum are connected by lines across the years in which the stratum was fished. The bar represents twice the average of standard errors of the estimates.

All the models described were fitted using S-plus and further details are given in Candy (2003a).

Discussion

The results shown in Figures 8 and 9 are the most important from a fisheries management perspective since they represent the goal of model fitting and regression diagnostic analyses used to support the particular models employed. The key point to be noted from Figure 8 is that for the early years of the fishery the estimated CPUE is much less precise and is, in general, considerably greater for the glmm(link = log) fit compared to that for the glm(link = sqrt) fit. The glm(link = log) fit gives quite similar estimates to the glmm(link = log) but with much more 'apparent' precision. However, given the significant between-vessel variation in catch rates as demonstrated by Figure 4, it is clear

that single-error models are inadequate to model the structure in the true variability in the catch data conditional on effort and the predictive factors of year, nationality and depth class.

Figures 6 and 7 demonstrate empirically that a Tweedie distribution for catch weight adequately describes the distribution of estimated residuals for the model of catch weight, whereas the traditionally used model fails quite dramatically to adequately model the distribution of residuals for the model of haul-by-haul CPUE. However, there are limitations to the Tweedie or CPG distribution as a model of haul-by-haul (i.e. aggregated) catch weight. For example, a single gamma distribution for individual fish weight is only an approximation when the population is a mixture of a number of year classes, each with a different weight distribution. Also, number of fish caught may be more dispersed than a Poisson distribution. Improvements in the ability of the CPG distribution to model catch weight may be possible using additional random effects, such as spatial random effects, to account for over-dispersion in the Poisson counts. This could be tested directly if counts of fish were available for each haul. Unfortunately, comprehensive catch counts are rarely available for commercial catch-and-effort data and without the benefit of such data the Tweedie or CPG distribution appears to be the best theoretical and empirical distribution for catch weight.

Punt et al. (2000) and Brandão et al. (2002) incorporated vessels as fixed effects and then standardised the CPUE series by either using the median of the estimated vessel coefficients (Brandão et al., 2002) or the vessel with the most records (i.e. hauls) (Punt et al., 2000). There are a number of drawbacks to this approach relative to the mixed-model approach adopted here. First and most serious is the problem of imbalance in the data which is greatly increased by including vessels as fixed effects. When this approach was tried for the longline data using model (2), with a 'sum-to-zero' parameterisation in S-plus for the fixed SHIP effect and dropping the nationality factor (since a fixed-effect vessel term renders the nationality fixed-effect term redundant), two vessels were completely confounded with YEAR, with S-plus giving missing values for their coefficients. After setting the coefficients for these two vessels to zero, the median of the vessel coefficient estimates was –0.08911. When the standardised CPUE series was calculated without scaling by exp(–0.08911) = 0.915, the estimates for seasons after 1991 were similar to those in Figure 8. However, for seasons prior to 1992 this approach gave estimates, with or without scaling, that were far too small (i.e. unscaled values of 0.2764, 0.0048, 0.0024 and 0.0033 for seasons 1987, 1988, 1989 and 1991 respectively). This poor estimation of standardised CPUE for these years can be attributed to the high degree of imbalance of vessels across the early years of the fishery and the effect this can have on 'treatment' (i.e. YEAR) fixed-effect parameter estimates obtained from the 'intra-block' analysis (i.e. vessels = 'blocks' fitted as fixed effects) (Robinson, 1991). For example, the 1987 season was fished by a single vessel. The mixed-model approach with vessels as random effects does not suffer from this problem since fixed effects are estimated by generalised least squares (GLS) (Patterson and Thompson, 1971) while, separately, BLUP estimation of random effects and REML estimation of the random-effect variance are interleaved in the iterative fitting algorithm (Laird and Ware, 1982). Fitting the GLMM provides more efficient estimates of the standardised CPUE time series. Extra efficiency is gained by recovery of inter-vessel information as a result of the imbalance in the number of hauls in the year by vessel cross-classification. In addition, fitting vessels that are by nature random (i.e. they 'come and go' within the fishing fleet) as fixed effects, will give biased estimates of the variance of the marginal estimates of standardised CPUEs, possibly seriously underestimating these variances if the vessel random-effect variance is large, as was the case for the longline data.

The key point to be noted from Figure 9 is that there was an increase in CPUE in the second year of the trawl fishery, with a sharp decline in the following year followed by a relatively stable series of CPUEs. It can also be seen that the influence on the series of the random-effects estimates for stratum-by-year is small (i.e. the most extreme *SxY*_re estimate in Figure 10(b) and 12 of –0.844 corresponds in Figure 9 to year 2001 in Stratum 2).

Figures 10 and 11 show that for both sets of estimated random effects their distributions are reasonably well approximated by a Gaussian distribution, though the stratum-by-year effects tend to have a greater kurtosis (2.3, s.e. = 0.5) than expected for a Gaussian distribution.

Fitting the stratum-by-year interaction as a fixed effect in an LM or GLM (e.g. Punt et al., 2000; Brandão et al., 2002) requires the areal extent of each stratum to be known in order to 'average out' stratum to obtain standardised yearly CPUE estimates for the overall fishery. Using the formula based on stratified random sampling to do this is problematic, given that hauls from commercial fishing are not a spatially random sample within each stratum. Also, 'messy' methods of handling

missing combinations of stratum and year in the data are required. In contrast, 'averaging out' a random stratum-by-year interaction simply involves ignoring this term when forming standardised CPUE estimates. In addition, including the stratum term as a fixed effect simply results in uniform scaling of the CPUE series as is the case with other additive fixed effect terms in the GLMM. Figures 10 to 12 demonstrate that for the trawl data, stratum-by-year effects can reasonably be considered as random (i.e. the probability-like distribution seen in Figures 10(b) and 11 and the general lack of obvious trends over years within strata or groupings of strata in Figure 12). Therefore separate standardised CPUE series for each stratum do not need to be calculated and then combined using an area weighting system, with all its uncertainties.

Brandão et al. (2002) also included interactions of year with other fixed effect terms in addition to stratum (='area') in a linear model, specifically month-by-year and vessel-by-year. The method of dealing with a year interaction described above for stratum-by-year can also be applied to other 'fixed term'-by-year interactions, thus simplifying the calculation of the standardised CPUE series.

Conclusions

From the model diagnostics presented for Subarea 48.3 it is clear that the GLMM with catch as the response variable, a log link, log of effort as an offset, random vessel effects in the intercept, and a (conditional) variance power function of 1.3 is much superior, particularly in terms of the goodness-of-fit of the link function (Figure 4), to the standard model fitted with haul-level C/E values as the response variable combined with a square-root link function and unit variance function.

Fitting a variance power function with power parameter estimated by the extended quasi-deviance criterion has allowed an intuitively sensible model of catch distribution, conditional on fixed and random effects, to be used in the form of a Tweedie distribution. The Tweedie distribution allows zero catches to be sensibly incorporated with non-zero catches in a single modelling procedure, and is well suited to modelling catch data (Figures 5 to 7) while clearly the assumption of a normal distribution for the standard model (1) residuals is not supported by the diagnostic plots presented (Figures 1 and 7).

When a stratum-by-year interaction behaves like a random variable it is a much simpler and more robust approach to include it as a random effect in a GLMM.

Acknowledgements

The author is grateful for Dr David Ramm's assistance in obtaining the longline data and for help by Tim Lamb and Dick Williams with the trawl data. The S-plus functions of Dr Gordon Smyth were used to fit and simulate the Tweedie distribution GLMs (see www.statsci.org/s/index.html). Thanks are also due to Drs Andrew Constable, Campbell Davies and Pavel Gasyukov and an anonymous referee for their helpful comments on the manuscript.

References

Brandão, A., D.S. Butterworth, B.P. Watkins and D.G.M. Miller. 2002. A first attempt at an assessment of the Patagonian toothfish (*Dissostichus eleginoides*) resource in the Prince Edwards Islands EEZ. *CCAMLR Science*, 9: 11–32.

Breslow, N.E. and D.G. Clayton. 1993. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88: 9–25.

Candy, S.G. 2000. The application of generalised linear mixed models to multi-level sampling for insect population monitoring. *Environmental and Ecological Statistics*, 7 (3): 217–238.

Candy, S.G. 2003a. Modelling catch and effort data using generalised linear models, the Tweedie distribution, and random vessel effects: longline fishery for *Dissostichus eleginoides* in CCAMLR Subarea 48.3. Document *WG-FSA-SAM-03/12*. CCAMLR, Hobart, Australia.

Candy, S.G. 2003b. Modelling catch and effort data using generalised linear models with random cruise and stratum-by-year effects: trawl fishery for *Dissostichus eleginoides* in CCAMLR Division 58.5.2. Document *WG-FSA-03/34*. CCAMLR, Hobart, Australia.

Diggle, P.J., K.-Y. Liang and S.L. Zeger. 1994. *Analysis of Longitudinal Data*. Clarendon Press, London.

Grömping, U. 1996. A note on fitting a marginal model to mixed effects log-linear regression data via GEE. *Biometrics*, 52: 280–285.

Jørgensen, B. 1997. *Theory of Dispersion Models*. Chapman and Hall, London: Chapter 4.

Laird, N.M. and J.H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics*, 38: 963–974.

Lee, Y., and J.A. Nelder. 2001. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88: 987–1006.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models.* 2nd Edition. Chapman and Hall/CRC, USA.

Patterson, H.D. and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58: 545–554.

Punt, A.E., T.I. Walker, B.L. Taylor and F. Pribac. 2000. Standardization of catch and effort data in a spatially-structured shark fishery. *Fish. Res.*, 45 (2): 129–145.

Robinson, G.K. 1991. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6: 15–51.

SC-CAMLR. 2002. WG-FSA Standard Assessment Methods. Document *SC-CAMLR-XXI/BG/28*. CCAMLR Hobart, Australia: 25 pp.

Schall, R. 1991. Estimation in generalised linear models with random effects. *Biometrika*, 78: 719–27.

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

Smyth, G. K. and A.P. Verbyla. 1999. Adjusted likelihood methods for modelling dispersion in generalised linear models. *Environmetrics*, 10: 695–709.

Zeger, S.L., K.-Y. Liang and P.S. Albert. 1988. Models for longitudinal data: generalised estimating equation approach. *Biometrics*, 44: 1049–1060.

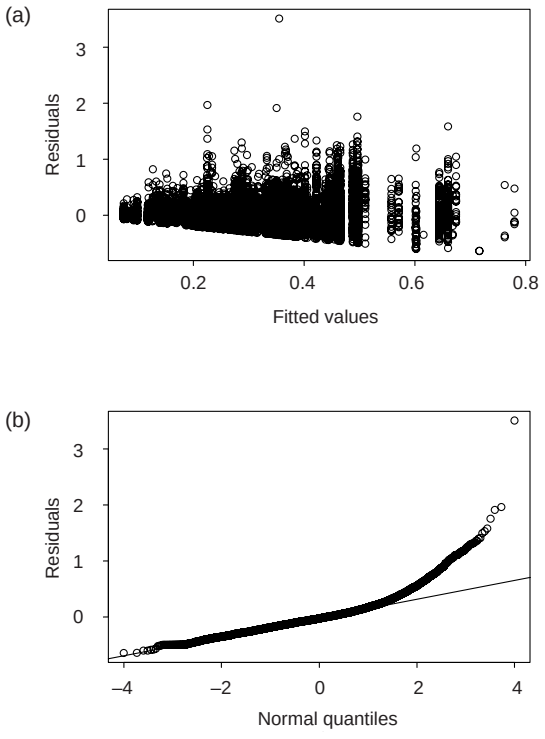

Figure 1: Ordinary residuals from the fit of the glm(link = sqrt) with C/E as the response variable: (a) residuals versus fitted values, (b) normal QQ plot.

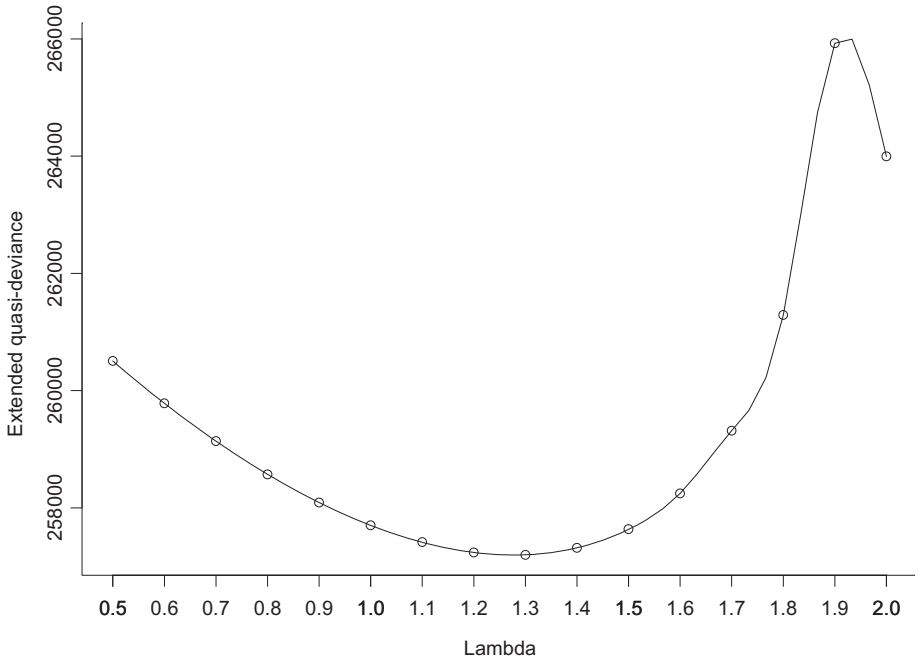

Figure 2: Profile extended quasi-deviance for lambda parameter in the Tweedie GLMM with log link.

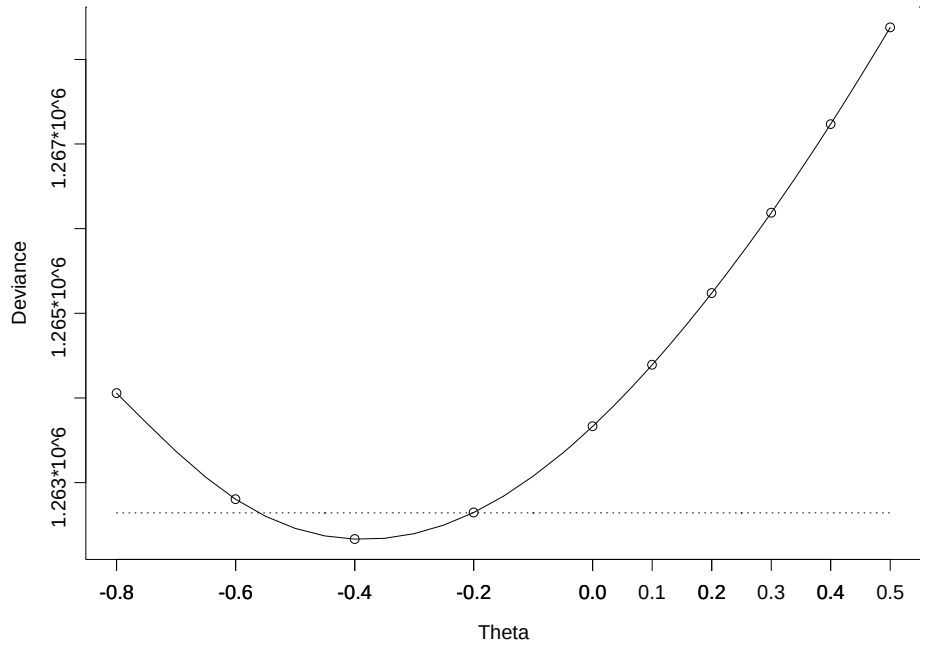

Figure 3: Profile deviance (Lambda = 1.3) versus link power parameter, Theta. Log and sqrt link functions correspond to Theta = 0, 0.5 respectively. The interval with endpoints given by the intersection of the dotted line with the profile represents an approximate 95% support interval for Theta.

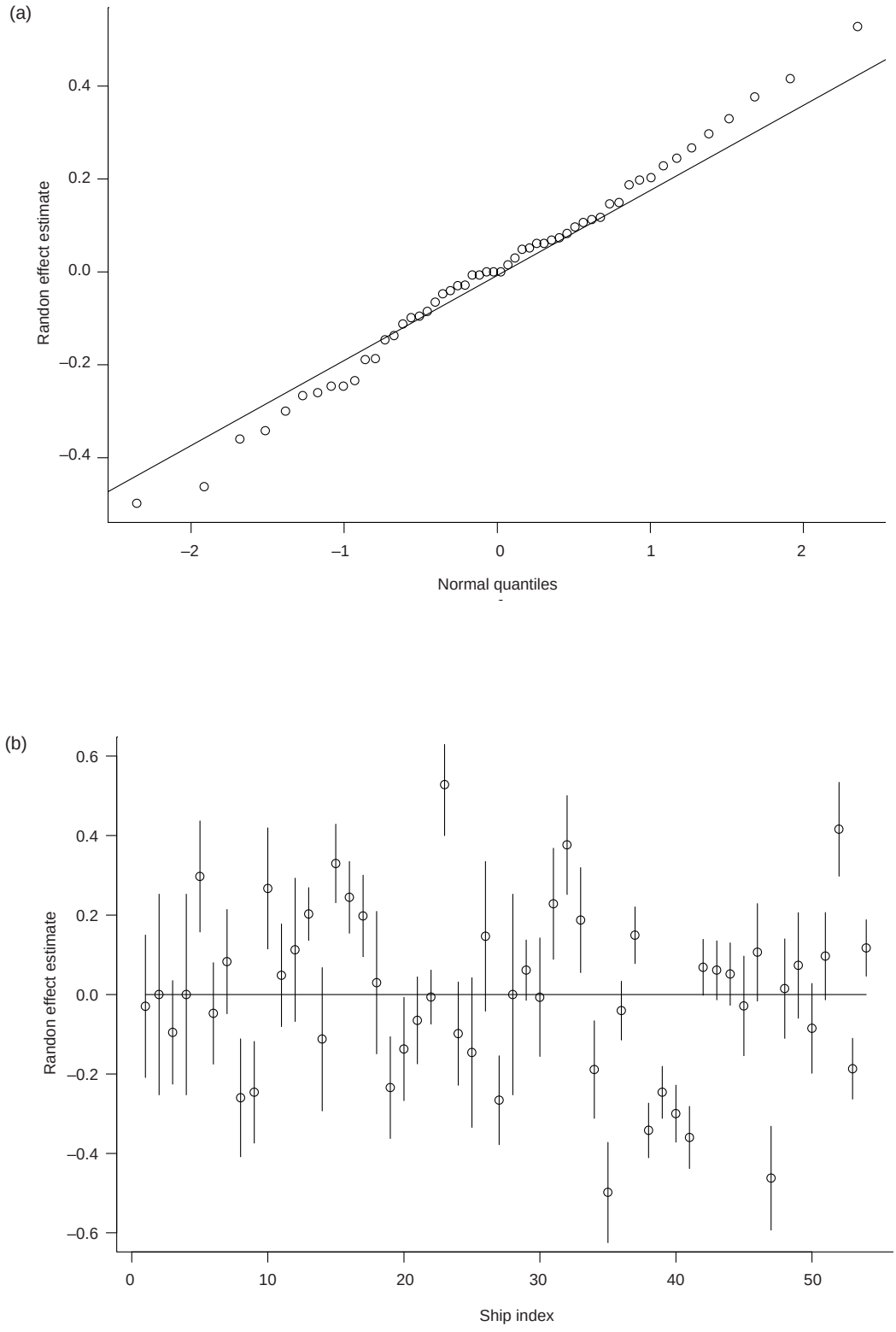

Figure 4: Random SHIP effect estimates from the fit of the GLMM with log link, and lambda = 1.3 to catch values after removing outliers: (a) normal QQ plot, (b) random-effect estimates versus SHIP index (1 to 54) with standard error bars shown.

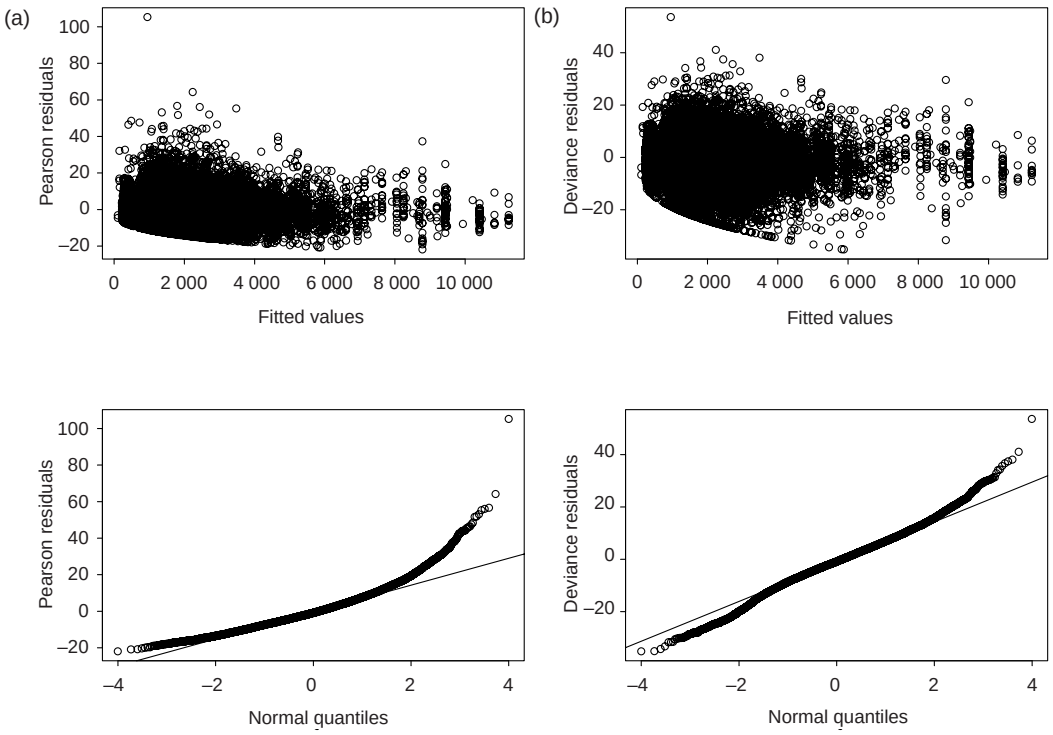

Figure 5: Conditional residuals from the fit of the GLMM (log link) with lambda = 1.3. (a) Pearson residuals versus fitted values and the corresponding normal QQ plot, (b) deviance residuals versus fitted values and the corresponding normal QQ plot.

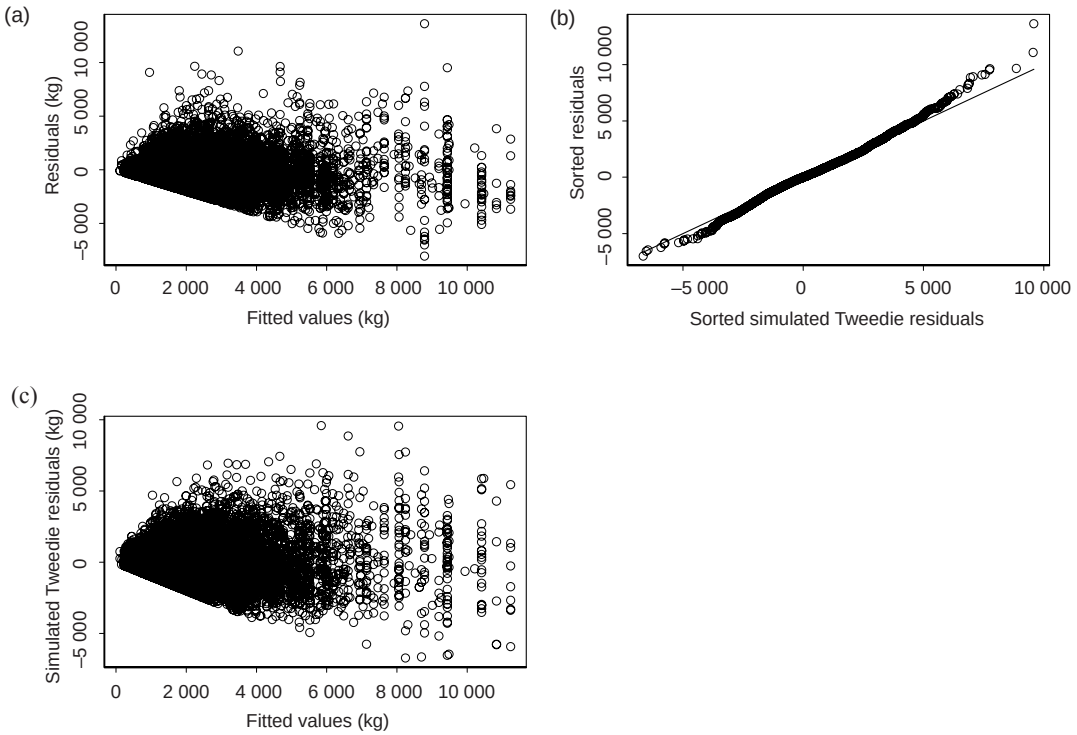

Figure 6: Conditional ordinary residuals from the fit of the GLMM (log link) with lambda = 1.3. (a) Observed residuals versus fitted values, (b) simulated Tweedie distribution residuals, (c) quantiles of observed and simulated catches using fitted values and lambda = 1.3 and phi = 68.4.

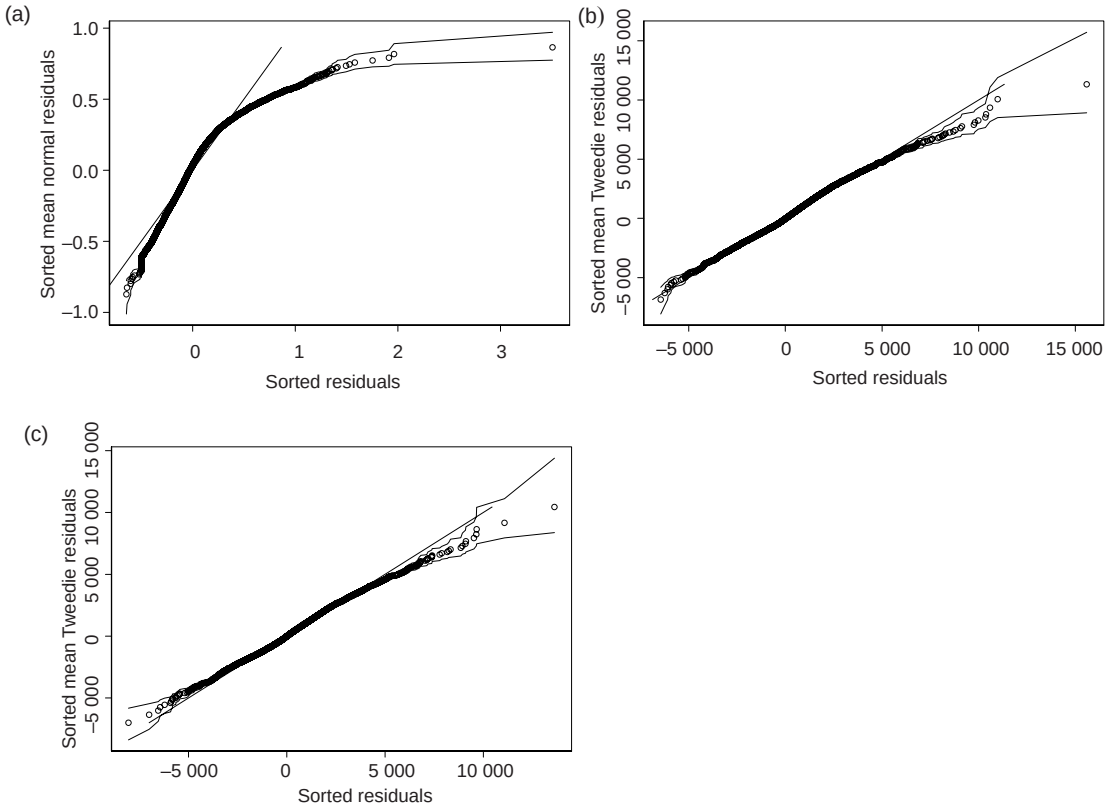

Figure 7: Sorted ordinary observed residuals from the fit of: (a) glm(square root) to CPUE haul values, (b) glm(log), (c) glmm(log) to catch values versus mean of simulated ordinary residuals for normal or Tweedie distributions for mean CPUE or catch given by the model fitted values. The residuals for (c) are conditional on the estimated random-vessel effects and the envelope represents a 96% confidence bound from 100 simulations for each of the 15 712 fitted values.

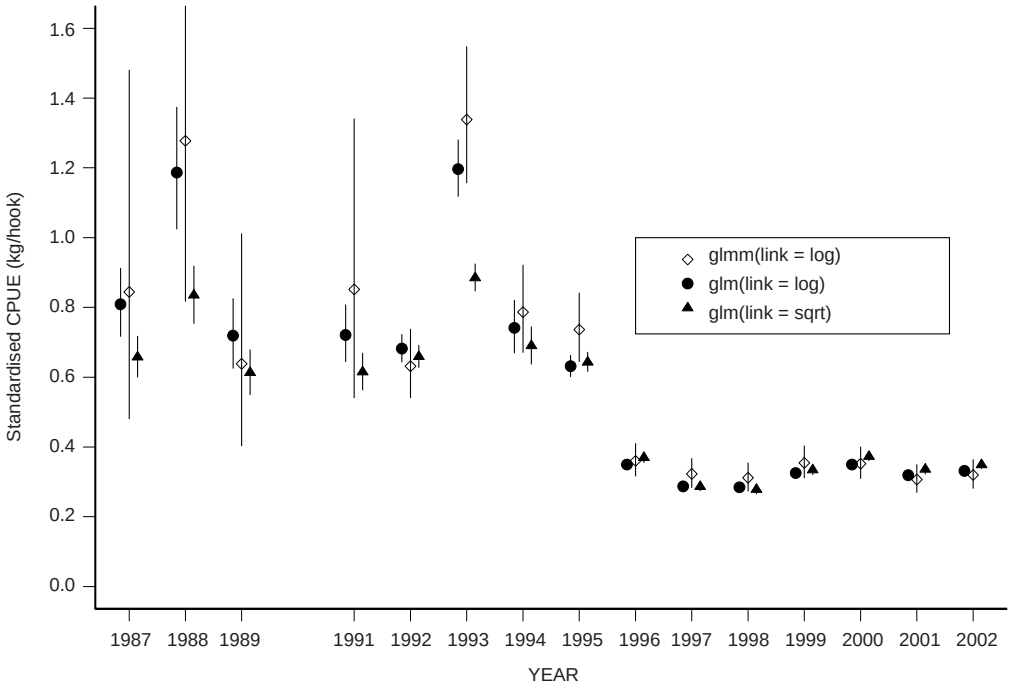

Figure 8: Standardised CPUE by fishing year estimated from each of three models. The variance power parameter was 1.3 for the glm(link = log) and glmm models. The square-root link model was fitted to the C/E haul data as response while the log-link model used C as response with $\log(E)$ as an offset. Approximate 95% confidence bars are shown and the estimates for each model are for the same year but shown with offsets on the YEAR axis to improve clarity.

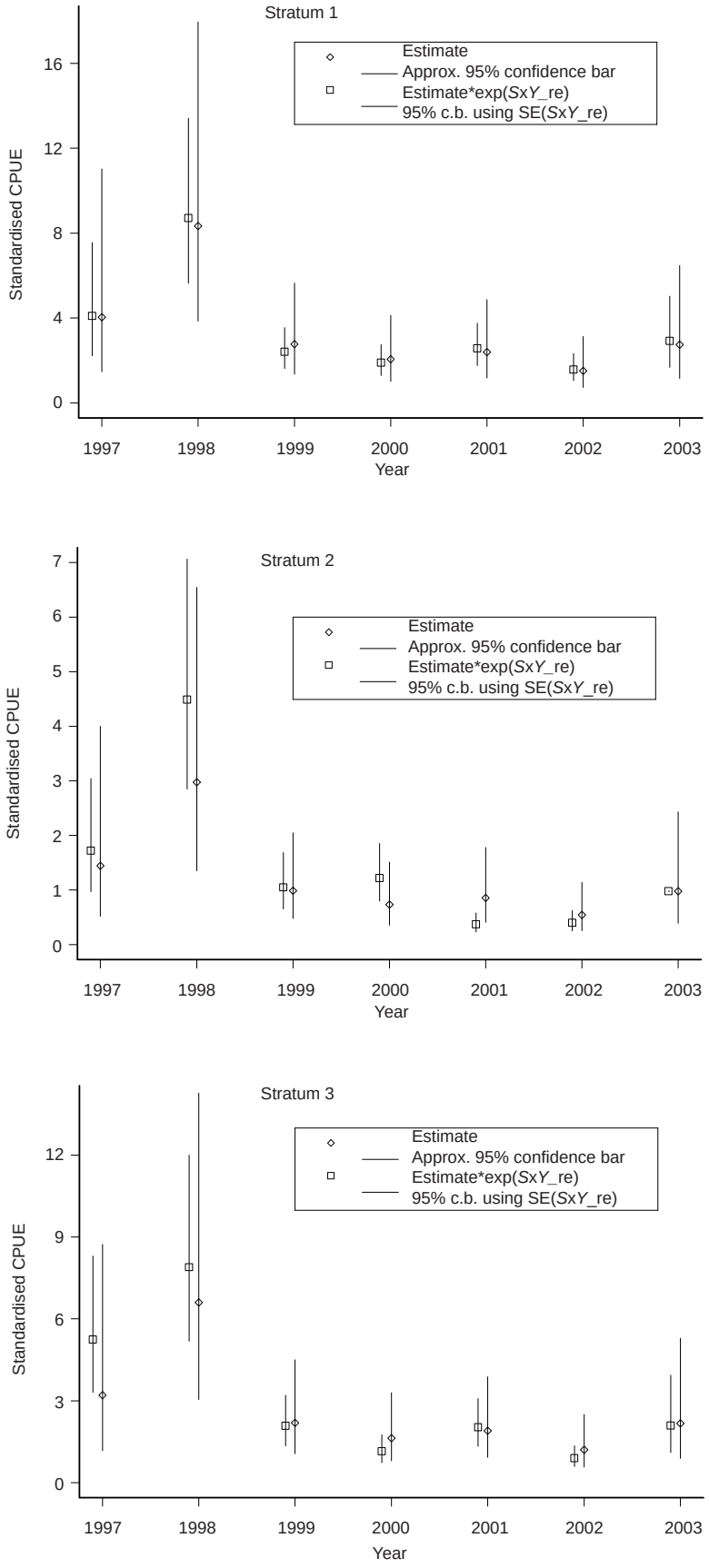

Figure 9: Standardised CPUE by fishing season for three strata with estimates obtained from the GLMM with random cruise and stratum-by-year (*SxY*) effects. The parallel series with random *SxY* effects included in estimates are also shown with approximate 95% confidence bars calculated as Estimate*exp{*SxY*_re+/-1.96*SE(*SxY*_re)}.

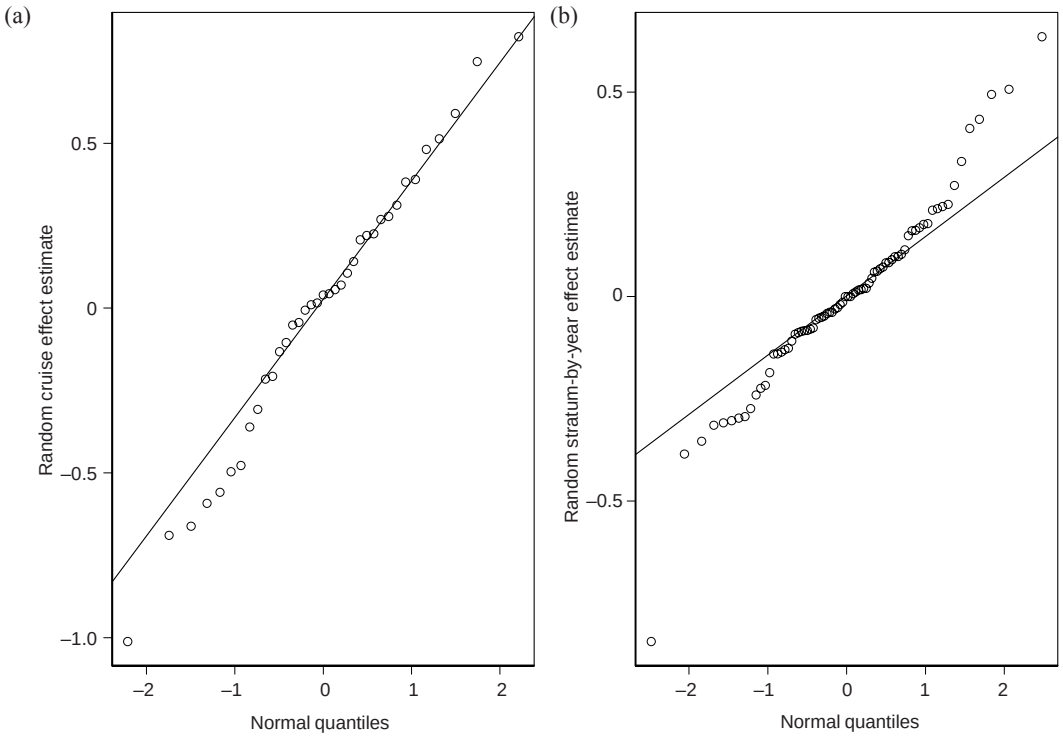

Figure 10: Normal quantile–quantile plots for estimated random effects from the fit of the GLMM: (a) random cruise effect estimates, (b) random stratum-by-year effect estimates.

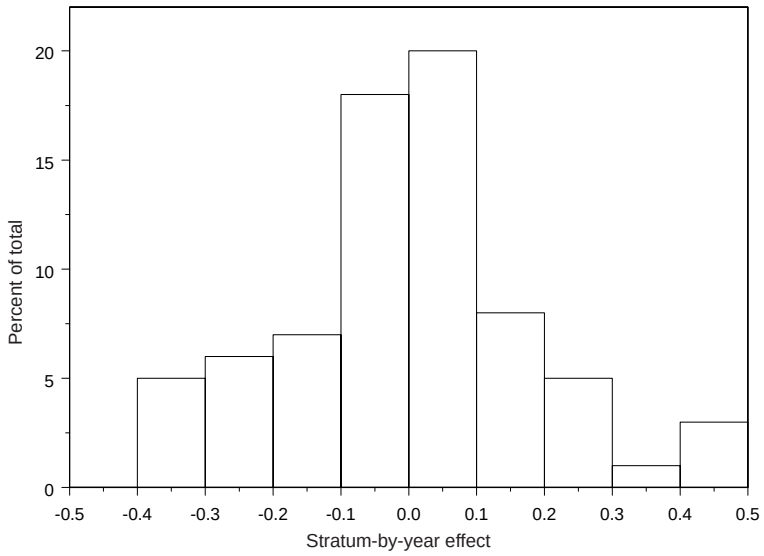

Figure 11: Frequency histogram of random stratum-by-year effect estimates (*SxY*_re).

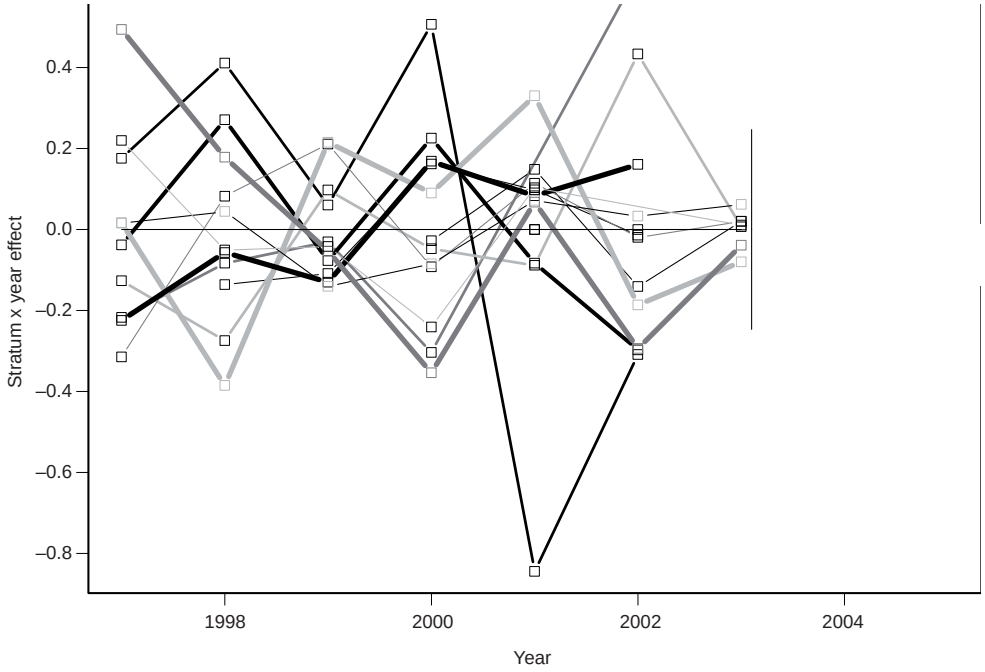

Figure 12: Stratum-by-year random-effect estimates (*SxY*_re) from the fit of the GLMM with log link and lambda = 1.6. Estimates for the same stratum are connected by lines across the years in which the stratum was fished. The bar represents twice the average of standard errors of the estimates.

Liste des figures

Figure 1: Résidus ordinaires de l'ajustement du glm(lien = racine carrée) avec C/E comme variable réponse : (a) résidus contre valeurs ajustées, (b) QQ plot normal.

Figure 2: Quasi-déviance du profil étendue du paramètre de lambda dans le GLMM de Tweedie avec lien log.

Figure 3: Déviance du profil (Lambda = 1,3) contre le paramètre de puissance du lien, Theta. Les fonctions liens Log et racine carrée correspondent respectivement à Theta = 0, 0,5. L'intervalle dont les points d'extrémité sont donnés par l'intersection de la ligne en pointillés et du profil représente un intervalle de confiance à environ 95% pour Theta.

Figure 4: Estimations de l'effet aléatoire SHIP de l'ajustement du GLMM avec le lien log, et lambda = 1,3 aux valeurs des captures après la suppression des aberrations : (a) QQ plot normal, (b) estimations des effets aléatoires contre indice SHIP (1 à 54) avec indication des barres d'erreur standard.

Figure 5: Résidus conditionnels de l'ajustement du GLMM (lien log) avec lambda = 1,3. (a) résidus de Pearson contre les valeurs ajustées et le QQ plot normal correspondant, (b) résidus de la déviance contre les valeurs ajustées et le QQ plot normal correspondant.

Figure 6: Résidus conditionnels ordinaires de l'ajustement du GLMM (lien log) avec lambda = 1,3. (a) résidus observés contre valeurs ajustées, (b) résidus simulés de la distribution de Tweedie, (c) quantiles des captures observées et simulées au moyen de valeurs ajustées et de lambda = 1,3 et phi = 68,4.

Figure 7: Résidus ordinaires observés et triés de l'ajustement de : (a) glm(racine carrée) aux valeurs de la CPUE par pose, (b) glm(log), (c) glmm(log) aux valeurs de capture contre la moyenne des résidus ordinaires simulés des distributions normales ou de Tweedie de la CPUE moyenne ou capture donnée par les valeurs ajustées du modèle. Les résidus de (c) dépendent des effets aléatoires estimés des navires; l'enveloppe représente un intervalle de confiance à 96% pour 100 simulations de chacune des 15 712 valeurs ajustées.

Figure 8: CPUE normalisée par année de pêche estimée par chacun des trois modèles. Le paramètre de puissance de la variance était de 1,3 pour le glm(lien = log) et le glmm. Le modèle lien racine carrée a été ajusté aux données de pose C/E en tant que réponse, alors que le modèle lien log a utilisé C comme réponse avec log(E) pour compenser. Des barres de confiance à environ 95% sont indiquées; les estimations de chaque modèle sont pour la même année mais sont compensées sur l'axe des YEAR pour plus de clarté.

Figure 9: CPUE normalisée par saison de pêche pour trois strates; les estimations proviennent du GLMM avec effets aléatoires des campagnes et des strates selon l'année (SxY). Les séries parallèles avec les effets aléatoires SxY compris dans les estimations sont également indiquées avec des barres de confiance à environ 95% calculées comme suit, Estimate*exp{SxY_re+/-1,96*SE(SxY_re)}.

Figure 10: Tracés quantile–quantile pour estimer les effets aléatoires de l'ajustement du GLMM: (a) estimations des effets aléatoires des campagnes, (b) estimations des effets aléatoires de la strate selon l'année.

Figure 11: Histogramme de fréquence des estimations des effets aléatoires des strates selon l'année (SxY_re).

Figure 12: Estimations des effets aléatoires des strates selon l'année (SxY_re) de l'ajustement du GLMM avec le lien log et lambda = 1,6. Les estimations d'une même strate sont connectées par des lignes les années où elle a fait l'objet de pêche. La barre représente le double de la moyenne des erreurs standard des estimations.

Список рисунков

Рис. 1: Нормальные остатки при подборе glm(связь = корень квадратный), где – C/E зависимая переменная: (a) остатки по сравнению с подобранными значениями, (b) нормальный QQ график.

Рис. 2: Квазилинейное отклонение с расширенным профилем для параметра лямбда в GLMM Твиди с логарифической связью.

Рис. 3: Отклонение профиля (лямбда = 1.3) по отношению к показателю степени связи – тэта. Функции связи логарифмическая и в виде корня квадратного соответственно относятся к тэта = 0, 0.5. Интервал с конечными точками, полученными при пересечении пунктирной линии с профилем, является приблизительным 95%-ным опорным интервалом для тэта.

Рис. 4: Оценки случайного эффекта SHIP из GLMM с логарифмической связью, при лямбда = 1.3 для получения значений после удаления выбросов: (a) нормальный QQ график, (b) оценки случайных эффектов по отношению к индексу SHIP (1–54), показаны величины стандартных ошибок.

Рис. 5: Условные остатки при подборке GLMM (логарифмическая связь связь) при лямбда = 1.3. (a) остатки Пирсона по отношению к подобранным значениям и соответствующий этому нормальный QQ график, (b) остатки отклонений по отношению к подобранным значениям и соответствующий нормальный QQ график.

Рис. 6: Условные нормальные остатки при подборе GLMM (логарифмическая связь) при лямбда = 1.3. (a) наблюдаемые остатки по отношению к подобранным значениям, (b) смоделированные остатки распределения Твиди, (c) квантили наблюдаемых и смоделированных уловов с использованием подобранных значений и при лямбда = 1.3 и фи = 68.4.

Рис. 7: Распределенные нормальные наблюдаемые остатки при подборе: (a) glm(с корнем квадратным) к значениям CPUE выборки, (b) glm(log), (c) glmm(log) к значениям улова по отношению к среднему смоделированных нормальных остатков при нормальном распределении или распределении Твиди для среднего CPUE или улова, полученного по подобранным значениям. Остатки в (c) зависят от оценки случайных эффектов по судам, а огибающая кривая представляет собой 96%-ную доверительную границу, полученную в результате 100 расчетов для каждого из 15 712 подобранных значений.

Рис. 8: Стандартизованные CPUE по годам промысла, оцененные по каждой из трех моделей. Показатель степени дисперсии составил 1.3 для glm(связь = log) и glmm. В модели со связью в виде квадратного корня в качестве зависимой переменной использовались данные по выборке C/E, тогда как в модели с логарифмической связью зависимой переменной было C, а лог (E) – смещение. Показаны приближенные 95%-ные доверительные пределы; оценки по каждой модели даются за один и тот же год, но для большей четкости показаны со смещениями по оси ГОД (YEAR).

Рис. 9: Стандартизованные CPUE за промысловый сезон по трем зонам с оценками, полученными по GLMM со случайными эффектами для рейсов и зон по годам (*S*x*Y*). Также показаны параллельные ряды с включенными в оценки случайными эффектами *S*x*Y* с приближенными 95%-ными доверительными пределами, рассчитанными как Estimate*exp{*S*x*Y*_re+/-1.96*SE(*S*x*Y*_re)}.

Рис. 10: Нормальные квантиль-квантиль графики для рассчитанных случайных эффектов по GLMM: (a) оценки случайных эффектов для рейсов, (b) оценки случайных эффектов для зон по годам.

Рис. 11: Частотная гистограмма оценок случайных эффектов случайных зон по годам (*S*x*Y*_re).

Рис. 12: Оценки эффектов случайных зон по годам (*S*x*Y*_re) по GLMM с логарифмической связью и лямбдой = 1.6. Оценки для одной и той же зоны соединены линиями через те годы, когда в данной зоне велся промысел. Прямоугольник представляет собой удвоенное среднее для стандартных ошибок оценок.

Lista de las figuras

Figura 1: Residuales ordinarios del ajuste del glm(enlace = sqrt) con C/E como la variable de respuesta: (a) residuales *vs* valores ajustados, (b) gráfico q-q normal.

Figura 2: Cuasi-desvianza de perfil extendido del parámetro lambda en el GLMM de Tweedie con enlace logarítmico.

Figura 3: Desvianza del perfil (Lambda = 1.3) *vs* el parámetro de potencia de enlace, Theta. Las funciones logarítmicas y de raíz cuadrada corresponden a Theta = 0, 0.5 respectivamente. El intervalo cuyos puntos extremos están dados por la intersección de la línea punteada con el perfil representa un intervalo de apoyo de un 95% aproximadamente para Theta.

Figura 4: Las estimaciones del efecto aleatorio SHIP derivadas del ajuste del modelo GLMM con enlace logarítmico, y lambda = 1.3 para los valores de captura después de eliminar los valores atípicos: (a) gráfico q-q normal, (b) se muestran las estimaciones del efecto aleatorio *vs* el índice SHIP (1 a 54) con las barras del error estándar.

Figura 5: Residuales condicionales del ajuste del GLMM (enlace logarítmico) con lambda = 1.3. (a) residuales de Pearson *vs* valores ajustados y el gráfico q-q normal correspondiente, (b) residuales de la desvianza *vs* los valores ajustados y el gráfico q-q normal correspondiente.

Figura 6: Residuales ordinarios condicionales del ajuste del GLMM (enlace logarítmico) con lambda = 1.3. (a) residuales observados *vs* valores ajustados, (b) residuales simulados de la distribución Tweedie, (c) cuantilos de las capturas observadas y simuladas mediante los valores ajustados y lambda = 1.3 y phi = 68.4.

Figura 7: Clasificación de los residuales ordinarios observados del ajuste de: (a) glm(raíz cuadrada) a los valores CPUE del lance, (b) glm(log), (c) glmm(log) a los valores de captura *vs* el promedio de los residuales ordinarios simulados para las distribuciones normal o de Tweedie del CPUE promedio o la captura obtenida de los valores ajustados al modelo. Los residuales para (c) dependen de la estimación de los efectos aleatorios del factor barco y la envoltura representa un intervalo de confianza del 96% de 100 simulaciones para cada uno de los 15 712 valores ajustados.

Figura 8: CPUE normalizado por año de pesca estimado de cada uno de los tres modelos. El valor del parámetro de potencia para la varianza fue 1.3 para los modelos glm(enlace logarítmico) y glmm. El modelo con una función de enlace de raíz cuadrada fue ajustado a los datos de lances C/E como respuesta, mientras que el modelo con enlace logarítmico utilizó C como respuesta, compensando con log(E). Se muestran los intervalos de confianza aproximados del 95%. Las estimaciones para cada modelo corresponden al mismo año pero se muestran compensadas en el eje YEAR para aumentar la claridad.

Figura 9: CPUE normalizado por temporada de pesca para tres estratos con estimaciones derivadas del GLMM con efectos aleatorios correspondientes a la campaña y estrato por año (*S*x*Y*). También se muestran las series paralelas con efectos aleatorios *S*x*Y* incluidos en las estimaciones con intervalos de confianza aproximados de 95%, calculados como Estimate*exp{*S*x*Y*_re+/-1.96*SE(*S*x*Y*_re)}.

Figura 10: Gráficos cuantilo-cuantilo normales para las estimaciones de los efectos aleatorios del ajuste del GLMM: (a) estimaciones del efecto aleatorio de la campaña, (b) estimaciones del efecto aleatorio estrato-año.

Figura 11: Histograma de la frecuencia de las estimaciones del efecto aleatorio del estrato-año (*S*x*Y*_re).

Figura 12: Estimaciones del efecto aleatorio del estrato-año (*S*x*Y*_re) del ajuste del GLMM con enlace logarítmico y lambda = 1.6. Las estimaciones para el mismo estrato están conectadas por líneas a través de los años en que el estrato fue explotado. La barra representa el doble del promedio del error estándar de las estimaciones.

APPENDIX 1

DENSITY FUNCTION OF THE COMPOUND POISSON OR TWEEDIE DISTRIBUTION

Following Smyth (1996), the probability-density function (PDF) for Y where Y is a GLM response variable with power variance function, $V(\mu) = \mu^{\lambda}$, based on quasi-likelihood theory (McCullagh and Nelder, 1989), satisfies

$$\frac{\partial \log_e f_Y}{\partial \mu} = \frac{\partial}{\partial \mu}\int^{\mu} \frac{y - t}{\phi t^{\lambda}} dt = \frac{y - \mu}{\phi \mu^{\lambda}}.$$

At the same time, the PDF for a compound Poisson-gamma distribution for Z, where

$$Z = W_1 + W_2 + \ldots W_k + \ldots + W_N$$

where the W_k ($k = 1,\ldots,N$) are independently and identically distributed gamma variables with mean μ_w and variance $\phi_w \mu_w^2$, and N is distributed as a Poisson variable with mean τ, has the same cumulant-generating function as that of Y (Smyth, 1996) for $1 < \lambda < 2$. Therefore the PDF for Y can be expressed in terms of that for Z as

$$\begin{aligned} f(y;\mu,\phi,\lambda) &= P(N=0)\delta(y) + \sum_{r=1}^{\infty} P(N=r) f_{Z|N=r}(y) \\ &= e^{-\tau}\,\delta(y) + \sum_{r=1}^{\infty} \frac{\tau^r e^{-\tau}}{r!} \frac{1}{\Gamma\left(r\phi_w^{-1}\right)} \left(\frac{y}{r\phi_w \mu_w}\right)^{\frac{1}{r\phi_w}} \exp\left(-\frac{y}{r\phi_w \mu_w}\right) d\{\log_e(y)\} \end{aligned}$$

where $\delta(y)$ is the Dirac delta function at zero and the relationships between parameters of the compound Poisson distribution and those of the corresponding Tweedie GLM are given by

$$\tau = \frac{1}{\phi}\frac{\mu^{2-\lambda}}{2-\lambda},\ \phi_w = \frac{\lambda - 1}{2-\lambda},\ \mu_w = \phi\,(2-\lambda)\mu^{\lambda-1} \qquad 1 < \lambda < 2.$$

Note that the gamma distribution for W_k is specified above in terms of the mean and dispersion parameters in order to avoid specifying a location parameter, ξ, where $0 < \xi < W_k$.

APPENDIX 2

DEVIANCE FUNCTION FOR THE TWEEDIE MODEL

The formula for deviance contribution, d_{ij}, for the power variance function with $1 < \lambda < 2$ is given by

$$d_{ij} = 2\frac{y_{ij}\left(y_{ij}^{1-\lambda} - \hat{\mu}_{ij}^{1-\lambda}\right)}{1-\lambda} - 2\frac{\left(y_{ij}^{2-\lambda} - \hat{\mu}_{ij}^{2-\lambda}\right)}{2-\lambda}.$$