Environ Ecol Stat
DOI 10.1007/s10651-012-0233-0

A Poisson–Gamma model for analysis of ecological non-negative continuous data

Scott D. Foster · Mark V. Bravington

Received: 4 May 2012 / Revised: 8 October 2012

Abstract The statistical analysis of continuous data that is non-negative is a common task in quantitative ecology. An example, and our motivation, is the weight of a given fish species in a fish trawl. The analysis task is complicated by the occurrence of exactly zero observations. It makes many statistical methods for continuous data inappropriate. In this paper we propose a model that extends a Tweedie generalised linear model. The proposed model exploits the fact that a Tweedie distribution is equivalent to the distribution obtained by summing a Poisson number of gamma random variables. In the proposed model, both the number of gamma variates, and their average size, are modelled separately. The model has a composite link and has a flexible mean-variance relationship that can vary with covariates. We illustrate the model, and compare it to other models, using data from a fish trawl survey in south-east Australia.

Keywords Compound gamma · Compound poisson · Generalised linear models · Poisson mixture of gammas · Tweedie · Zeros

1 Introduction

Many ecological studies generate data that is continuous but non-negative; for example, the total weight of a particular fish species in a single sample (trawl). Such data are (1) identically zero if the species is not caught, or (2) a positive real number if the species is caught (combined weight of all individual fish caught). These data show

Handling Editor: Ashis SenGupta.

S. D. Foster (✉)· M. V. Bravington
CSIRO's Division of Mathematics, Informatics and Statistics, CSIRO's Wealth from Oceans Flagship,
GPO box 1538, Hobart, TAS 7001, Australia
e-mail: scott.foster@csiro.au

Published online: 22 December 2012

many of the same characteristics, at least superficially, to data from other application areas; rainfall, insurance claims, and failure times are examples. The purpose of analysis is to try to understand the relationships between one of these (outcome) variables and other (co-)variables—a regression-like analysis.

The analysis of fish surveys provide an excellent showcase for the breadth of statistical models used for non-negative data containing exact zeros. The models are typically based on commonly used statistical models, for example log-normal regression or gamma generalised linear models (GLMs; see McCullagh and Nelder 1989). We classify the methods into three distinct categories:

1. Methods that subset or transform the data to remove zeros, and then use standard models (e.g. Brynjarsdottir and Stefansson 2004).
2. Methods that treat the zero class as special and provide a separate model for them (e.g. Stefansson 1996; Punt et al. 2000; Ortiz and Arocha 2004).
3. Methods that employ a distribution that is defined on the non-negative real numbers but do not treat zero as a special case. The usual choice for this distribution is the Tweedie distribution (e.g. Candy 2004; Shono 2008; Tascheria et al. 2010; Peel et al. 2012), but also see Ancelet et al. (2010).

The first set of methods change the data before its formal inspection and may inadvertently alter the relationships in the data. These methods rely on the unstated assumption that the relationships will be the same with or without the zeros. Although this may be justifiable for some datasets, the assumption is not safe in general, and we do not consider such models further in this manuscript.

The second category of models have been used frequently, especially within fisheries research (delta models; see Maunder and Punt 2004 and references therein). It is easy to understand why—they are straightforward to implement and provide a flexible modelling framework. Delta models have many similarities to models used for zero-inflated models for count (abundance) data (e.g. Aitchison 1955; Mullahy 1986; Heilbron 1994; Ridout et al. 1998). However, there are drawbacks in terms of model structure; it is hard to impose a multiplicative structure.

A multiplicative structure on the expected value is often appropriate in ecological applications. A motivating example is the use of an offset term to reflect differing sampling effort. In fisheries data this corresponds to a trawl's swept-area, or trawl duration (a proxy for swept-area). A larger sampling effort should be reflected in a proportionally larger expectation. The same philosophy applies, for example, when trawls are taken at different depths in different regions. We expect that the ratio of catches between the depths for the different regions to be similar, even though the values of the catches may differ between regions. It is not difficult to envisage circumstances where these arguments do not strictly apply, but unless there is strong prior information or strong diagnostic indications to the contrary, they should be imposed.

The most common model employed in the third category, for fisheries applications at least, are Tweedie GLMs (Smyth 1996). The Tweedie GLM is not the only choice of stochastic distribution and others have been proposed (the law of leaks in Ancelet et al. 2010 for example), although this particular model can be shown to be closely related to a Tweedie with restrictions placed on the parameters. This category of models is appealing because no special treatment of zeros is easy to impose a multiplicative structure. Of course, there may be good reasons to treat zeros specially in particular situations—such as multiple generators for obtaining zeros (*sensu* Lambert 1992). However, these should not be the default (see Warton 2005).

We extend Tweedie GLMs by exploiting the compound Poisson formulation of the Tweedie and provide a model that has a flexible mean-variance relationship. This is achieved by allowing the parameters of each part of the Tweedie distribution's compound representation to vary as a function of covariates. The model is similar to that of Ancelet et al. (2010) except that we use a more flexible distribution and allow the mean and variance to vary with covariates.

The remainder of this document is organised as follows. In Sect. 2 commonly used methods are reviewed and our model is proposed. We present a maximum likelihood estimation method and explore the estimator's properties via simulation in Sect. 3. In Sect. 4, we describe a fish survey from south-eastern Australia. This data is an excellent resource for this purpose as, in addition to the biomass data, the researchers recorded the fish abundance (counts). The abundance data enables us to check the range of inferences of our proposed model. The results from the fish survey analysis are given in Sect. 4.1, which show the potential for our model. Finally, in Sect. 5 there is a summary and discussion. We provide software, in a web Appendix, that implements these methods.

2 Methods for non-negative continuous data

2.1 Delta log-normal model

The delta model analyses the entire data set without any subsetting. See Aitchison (1955) and Stefansson (1996) for a thorough description. Here we consider the delta log-normal model only, although any model defined on the positive real numbers can be used for non-zero data (e.g. Muralidharan and Kale 2002; Muralidharan and Lathika 2007). The delta log-normal model generalises log-normal regression by extending the support of the residuals to include zeros. This extension is achieved by defining the distribution as a mixture with probability π of observing a zero and $1-\pi$ of observing an observation from a log-normal distribution. Operationally, a delta log-normal model is fitted using a binary model fitted to the entire data, and a log-normal regression to the non-zero observations only. In this paper, we specify the mean structures for the ith observation, y_i, as

$$\begin{aligned}\text{logit}\left(\text{E}\left(\text{I}\left(y_i > \mathbf{0}\right)\right)\right) &= \boldsymbol{w}_i^{\top}\boldsymbol{\beta}, \text{ and}\\ \text{E}\left(\log(y_i)|y_i > \mathbf{0}\right) &= \boldsymbol{x}_i^{\top}\boldsymbol{\tau},\end{aligned}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$ are covariates for the ith observation, and $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are the associated parameters. If there are any zeros in the data then the design matrix for the log-linear model will have fewer rows than the design matrix for the binary model. The fitted value and the fitted variance from this model for the ith observation are (Aitchison 1955)

$$
\begin{aligned}
\mathrm{E}\left(y_i\right) &= \mathrm{E}\left(y_i > 0\right)\mathrm{E}\left(y_i | y_i > 0\right) \\
&= \operatorname{logit}^{-1}\left(\boldsymbol{w}_i^\top \boldsymbol{\beta}\right)\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\tau} + \frac{\sigma^2}{2}\right) \\
&= \pi_i \exp\left(\mu_i + \frac{\sigma^2}{2}\right) \text{ say, and} \\
\operatorname{Var}\left(y_i\right) &= \pi_i \exp\left(2\mu_i + \sigma^2\right)\left(\exp\left(\sigma^2\right) - \pi_i\right) \\
&= \left[\mathrm{E}\left(y_i\right)\right]^2\left(\frac{\exp\left(\sigma^2\right)}{\pi_i} - 1\right),
\end{aligned}
\tag{1}
$$

where σ^2 is the variance of the non-zero observations, on the log scale. The delta log-normal model does not provide a multiplicative model unless a log-link is used in place of the logit link. However, that choice of link can lead to invalid fitted values and predictions. We see the absence of a multiplicative model as a drawback for the delta models.

2.2 Tweedie generalised linear models

The Tweedie GLM (TGLM; Smyth 1996) is the GLM defined by the mean and variance structure $g(\mathrm{E}(\boldsymbol{y})) = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\tau}$ and $\operatorname{Var}(y_i) = \phi\mu_i^p$, where μ_i is the ith expectation. Choosing $g(\cdot)$ to be the logarithmic link is appealing for fisheries contexts as it defines a multiplicative model. The TGLM's mean and variance uniquely define the stochastic model in the GLM as a Tweedie (Jørgenson 1997) and so standard estimation methods for GLMs can be used (Smyth 1996). Estimation of the dispersion (ϕ) and power (p) parameters are less straightforward but can be estimated with numerical methods (e.g. Smyth 1996; Dunn and Smyth 2005). We estimate the location, dispersion and power parameters jointly from the full log-likelihood, maximised using a quasi-Newton optimiser. The Tweedie log-likelihood and its derivative are calculated in the spirit of the series expansion method given in Dunn and Smyth (2005), details are given in Appendix A.1.

The Tweedie model should appeal to many quantitative ecologists as the mean-variance relationship is explicitly that defined by Taylor's power law (1961), an empirical law that can be generated from a range of processes (e.g. Kendal 2004; Kendal and Jørgensen 2011). The power parameter in the mean variance relationship is typically, but not always, between 1 and 2 (Kendal 2004). One of the motivating processes is this: for $1 < p < 2$ a Tweedie random variable can be thought of as a Poisson sum of gamma variables. That is

$$
y_i = \sum_{j=1}^{N_i} w_{ij} \tag{2}
$$

where $N_i \sim \text{Poisson}(\lambda_i)$ and $w_{ij} \overset{\text{iid}}{\sim} \text{gamma}(\alpha, \beta_i)$. We call the w_{ij} the summands and N_i their frequency throughout this manuscript. For our application, this formulation is appealing as the total weight of a fish species in a sample is the sum of schools of fish, of which there are a random number; note that a ‘school’ may consist of a solitary fish. This formulation of the Tweedie model has not received much attention. One exception is Ancelet et al. (2010) who restricted the gamma variables to be exponential, creating the law-of-leaks model.

There is a relationship between the parameterisations (Smyth 1996; Dunn and Smyth 2005). Namely,

$$\lambda_i = \tfrac{1}{\phi}\tfrac{\mu_i^{2-p}}{2-p}, \quad \alpha = \tfrac{2-p}{p-1}, \quad \text{and } \beta_i = \phi(p-1)\mu_i^{p-1}. \tag{3}$$

This relationship is exploited for evaluating the Tweedie density function (Dunn and Smyth 2005) and provides a convenient mechanism from transferring from one parameterisation to the other.

2.3 The Poisson–Gamma approach

Our model is based on the Poisson-sum-of-gammas formulation of the Tweedie; see (2). Without covariates there is no difference in the distribution defined by the Poisson-sum-of-gammas formulation, $(\lambda_i, \alpha, \beta_i)$, and that defined by standard Tweedie formulation (μ_i, ϕ, p). However, differences do arise when covariates are allowed to affect the parameters.

We model the expectations $\mathrm{E}(N_i)$ *and* $\mathrm{E}\left(w_{ij}\right)$ separately, rather than the marginal expectation $\mathrm{E}(y_i)$ that is the focus of the Tweedie GLM. The joint modelling is enabled by defining a composite link (Thompson and Baker 1981). These two links combine to provide a simple multiplicative marginal model. The compound and marginal models are

$$\begin{aligned} \log\left(\mathrm{E}(N_i)\right) &= \boldsymbol{x}_i^\top \boldsymbol{\tau} + \log(\eta_i), \\ \log\left(\mathrm{E}\left(w_{ij}\right)\right) &= \boldsymbol{z}_i^\top \boldsymbol{\beta}, \text{ and} \\ \log\left(\mathrm{E}(y_i)\right) &= \log\left(\mathrm{E}(N_i)\right) + \log\left(\mathrm{E}\left(w_{ij}\right)\right) + \log(\eta_i) = \boldsymbol{x}_i^\top \boldsymbol{\tau} + \boldsymbol{z}_i^\top \boldsymbol{\beta} + \log(\eta_i), \end{aligned} \tag{4}$$

where η_i is an offset, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are vectors of covariates for the Poisson and gamma parts of the model respectively, and $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ are parameter vectors. The marginal variance for this composite model is

$$\mathrm{Var}(y_i) = \tfrac{\alpha+1}{\alpha}\mathrm{E}(N_i)\left[\mathrm{E}\left(w_{ij}\right)\right]^2 = \tfrac{\alpha+1}{\alpha}\mathrm{E}\left(w_{ij}\right)\mathrm{E}(y_i) \tag{5}$$

and varies linearly with the Poisson expectation and quadratically with the gamma expectation.

The set of covariates in the two different parts of the model may be equal but the data and purpose of the analysis should dictate this. We assume that the offset (e.g. duration of sample) affects the expected number of summands, rather than the expected weight of those summands. This is attractive in a space sampling context, where an increase in area sampled should result in an increase in encounters but not their weights. For example, fisheries trawl data frequently involves an offset of swept area, or trawl duration; it is natural to assume that this affects N_i and not w_{ij}.

Based on the marginal expectation, one might expect that the parameters within both parts of the model would be aliased for covariates that are shared. However, their individual effect can be separated via the variance of the observations. This is shown, via simulation, in Sect. 3.

2.4 Model diagnostics and model fit

We use randomised quantile residuals (Dunn and Smyth 1996) to graphically diagnose deficiencies in the fitted model. These residuals have a number of appealing qualities: their null distribution is known and they can be calculated for any parametric model. These residuals have already been used in many situations, including the Tweedie model (Smyth 1996; Tascheria et al. 2010; Peel et al. 2012).

For the delta models it is tempting to diagnose each of the two models separately on each version of the data. This is the easiest form of diagnostics as the software for fitting GLMs automatically produces the necessary quantities. We see two problems with this approach: (1) passing each of these diagnostic measures may not imply that the overall fit is sufficient as the interaction of the separate models is of utmost importance, and (2) diagnosing the binary model is usually extremely difficult. Inspecting the randomised quantile residuals for the full delta model removes both these difficulties.

Production of randomised quantile residuals requires calculation of the cumulative distribution function (CDF) of each datum under the fitted model. This function is defined for all parametric models. Randomisation is used to aid visual inspection for non-unique residuals, such as those obtained from zero observations with identical covariates. For identically zero observations, this is done by sampling a uniform variate from the interval $(0, \pi)$ where π is the probability of observing zero. The resulting set of (randomised) quantile residuals will be uniformly distributed, up to sampling variation of the parameters. Typically, these uniform residuals are further transformed using the inverse normal CDF.

We use the residuals as the primary method for determining the fits of competing models. This contrasts other researchers (e.g. Dick 2004) who try to distinguish between models using AIC alone. Automatic comparison of AIC is convenient but it is not sufficient; while it does inform about which model provides a better fit, it does not indicate if any of the models provide an adequate fit.

3 Estimation and a simulation study

Estimation for the Poisson–Gamma model is performed using maximum likelihood. We estimate all location and dispersion parameters from the full likelihood, even though this will cause some bias for the estimates of the dispersion parameters (vanishes with increasing sample size). This mirrors our approach to estimating Tweedie

GLMs, as mentioned in Sect. 2. The log-likelihood contribution from the i observation is given by (for example Smyth 1996)

$$\ell_i = \begin{cases} -\lambda_i & y_i = 0 \\ \frac{y_i}{\beta_i} - \lambda_i - \log y_i + \log W(y_i, \lambda_i, \alpha, \beta_i) & y_i > 0 \end{cases} \tag{6}$$

where

$$W(y_i, \lambda_i, \alpha, \beta_i) = \sum_{j=1}^{\infty} \frac{\lambda_i^j \left(\frac{y_i}{\beta_i}\right)^{j\alpha}}{j!\Gamma(j\alpha)}$$

is a generalised Bessel function. Evaluating the generalised Bessel function involves summing an infinite series. Fortunately, this can be accurately approximated using a finite sum (Dunn and Smyth 2005). For the applications here, the number of terms to sum has empirically proven to be small (typically less than 50) resulting in quick calculations of the log-likelihood contributions.

The log-likelihood is maximised using a quasi-Newton method (e.g. Nash and Sofer 1996). This method is descent based and provides super-linear convergence to the maximum. It requires the first derivative of the log-likelihood, but not the second derivative. The first derivative is found by a similar method to that used for the evaluation of the log-likelihood and involves an finite sum approximation to an infinite sum. Details of computational methods for evaluation of the density and derivatives are given in Appendix A.1. The estimation routine is adequately fast; estimation of the model for the fisheries survey data takes only ~0.3 s for 169 observations and 10 parameters, slightly less than for full maximum likelihood estimation of the comparable TGLM.

For computational convenience we choose to estimate $\log \alpha$ and not α directly, as this transformation removes the positivity constraint. The quasi-Newton algorithm is started with all parameters equal to zero.

3.1 Simulation study

It is useful to examine the properties of the parameter estimators via a simulation study. A simulation study will inform if there is significant bias in any of the parameter estimates; due either to model formulation, to using asymptotic results for finite samples, or from using maximum likelihood estimation for dispersion parameters. To this end, we repeatedly simulate data and analyse it using the PGM. The form of the model and data, is chosen to mirror that from the south-east fishery (SEF) data; see Sect. 4.

We specify 2 covariates, whose values are taken from the depth and along-coast variables from the SEF data. We also use the logarithm of the area swept variable as a scaling variable, i.e. as an offset in (4). We allow the mean number of summands to vary with both depth and along-coast covariates and we allow the mean summand weight to vary with depth. The linear predictors were all formed from orthogonal polynomials

Table 1 Summary of parameter estimates for the simulation study ($n = 10,000$)

Model part	Covariate	Term	Value	Estimate (±SE[a])	SD[b]	ASD[c](±SE)
Poisson	Depth	Intercept	−2.428	−2.448 (0.001)	0.146	0.144 (<0.001)
		Linear	−8.147	−8.440 (0.019)	1.933	1.869 (0.002)
		Quadratic	−4.132	−4.210 (0.018)	1.826	1.748 (0.001)
	Along-coast	Linear	3.161	3.271 (0.017)	1.667	1.626 (0.001)
		Quadratic	−1.797	−1.936 (0.017)	1.700	1.631 (0.001)
		Cubic	−1.387	−1.383 (0.017)	1.741	1.671 (0.001)
Gamma	Depth	Intercept	1.703	1.636 (0.002)	0.202	0.194 (<0.001)
		Linear	−11.668	−11.733 (0.026)	2.600	2.529 (0.004)
		Quadratic	−3.328	−3.747 (0.023)	2.338	2.249 (0.003)
Dispersion		$\log \alpha$	−0.755	−0.721 (0.002)	0.168	0.161 (<0.001)

[a] Empirical standard error of the mean of the 10,000 parameter estimates [b] Empirical standard deviation of the 10,000 parameter estimates [c] Mean asymptotic standard deviation of the parameter estimates (standard error)

of the original covariates. The order of the polynomials was 2 for frequency depending of depth, 3 for frequency depending on along-coast, and 2 for summand depending on depth. These parameters values are obtained by a Poisson–Gamma model fit to the real data from Sect. 4 (jackass morwong). The specific values are given in Table 1. We chose the number of observations as 169, to match the real data in Sect. 4.

The parameter estimates, in each of 10,000 replicates, were recorded along with their asymptotic variances and covariances. The results for the parameter estimates and estimates' variances (as standard deviations) are presented in Table 1. It is clear that the model's parameters are estimated well, although there is some slight bias in a few of the parameters. The dispersion parameter has very mild bias and is under-estimated, as is expected when estimating dispersions from full likelihoods.

The asymptotic standard deviations appear to reflect the true variation in the estimates. Again, there is some slight bias with the asymptotic variances tending to be slightly smaller than the empirical ones; see Table 1. For most practical purposes the level of this bias is likely to be inconsequential. The covariances (correlations) between the parameter estimates are modest—the range is (−0.25, 0.58). This implies that the separation of parameters for the frequency and summand models is adequate, at least for this simulation design.

We performed a similar simulation study for the Tweedie GLM estimation routine. This was done to make sure that the full maximum likelihood estimation method was efficacious. The results (see Appendix B) showed that the parameters were effectively estimated, although the slight bias is more noticeable.

4 South east fish survey data

The south east fishery (SEF) survey data are 169 fish trawl samples from an extensive fisheries-orientated survey (see Bax and Williams 1999, for details). The locations of

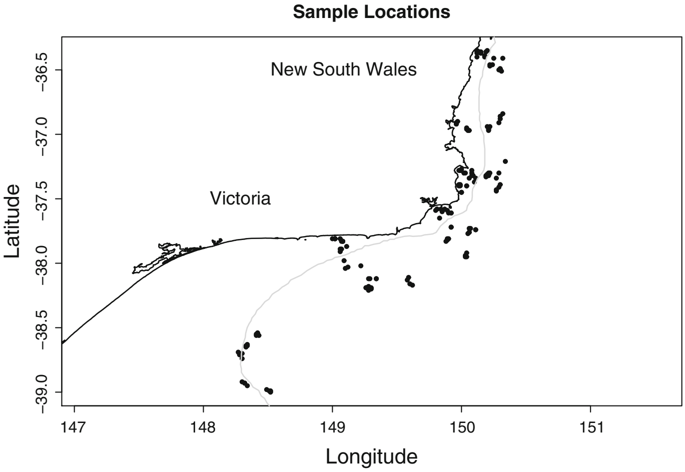

Fig. 1 Map of sample locations for the South-east Australia fishery data. *Black line* is the coast and *grey line* is approximately the 100 m depth contour. Note that the 100 m depth contour moves away from the coast in the south

the samples are shown in Fig. 1 and are representative of the mid to outer continental slope habitat in the region. We use data from 15 commercial species: pink ling, common jack mackerel, jackass morwong, john dory, redfish, gemfish, barracouta, deepsea flathead, tiger flathead, blue warehou, spotted warehou, bigeye ocean perch, silver trevally, eastern school whiting, and snapper. The species encountered most frequently was common jack mackerel (found at 148 sites) and the least encountered species was bigeye ocean perch (found at 13 sites).

One of the reasons for using the SEF data is that there were two types of measurements obtained on each species in each sample. These were: the total weight of all fish per species, and the number of fish per species (abundances). It is the availability of the abundance records that makes these data particularly useful for this study as they allow examination, in separate analyses, of the number of fish caught *and* their average weights.

We model species' catch weights on two variables: depth, and distance along coast (measured north to south along the 100m depth contour). These covariates do two things: (1) they effectively span the geographical space, and (2) provide surrogates for important biological drivers, such as temperature, salinity, oxygen, and so on. There is no reason to include an extra spatial dimension, such as distance from coast, as it would be highly confounded with depth.

The data for one example species, tiger flathead, is available in the R-package fishMod. The package is available from the Comprehensive R Archives Network (CRAN; http://cran.r-project.org/).

4.1 Results

We fitted four models to each species: a delta log-normal model (DLNM), a Tweedie GLM (TGLM), and two Poisson–Gamma models (PGM-1 and PGM-2). In all models, the dependence on covariates, if any, is specified using orthogonal polynomials in the linear predictor. The dependency on depth used a second order polynomial and the dependency on along-coast used a third order polynomial.

Both parts of the DLNM allowed for dependence on depth and along-coast. The log of the area swept variable was included as an offset in the log-normal model to account for varying trawl effort. The TGLM included dependencies on both covariates and an offset term for log area swept. Both PGMs included dependencies on both covariates in the Poisson part of the model. Also included in the Poisson part of the model, for both PGMs, was an offset for log area sampled. The two PGMs differ in their dependencies in the gamma part of the model. The first PGM (PGM-1) allowed the gamma part of the model to depend on depth whereas the second PGM (PGM-2) did not. Neither PGM-1 nor PGM-2 included dependence for distance along coast in the gamma part of the model.

We assess the appropriateness of the models via residual plots; see Fig. 2 for john dory and tiger flathead. Based solely on the assessment of fit from the residuls, we could find no reason to prefer any of the models. This is the case for all of the species modelled.

The fit of the two PGMs can be compared statistically. Such a comparison should reveal if the extra dependence on depth, through the depth terms in the gamma part of the model, is beneficial. The models are nested and so a formal likelihood ratio test is available. There was evidence that the extra depth effect is advantageous for jackass morwong ($p = 0.001$), spotted warehou ($p < 0.001$), and eastern school whiting ($p = 0.017$). There was also some further mild evidence that it was advantageous for barracouta ($p = 0.073$) and tiger flathead ($p = 0.063$). For the remaining species we would choose the simpler model for inference.

We now illustrate the fits of the different models using tiger flathead as an example. Model summaries from the different models are given in Fig. 3. All models provide qualitatively similar results; they all indicate that the catch rate of tiger flathead is maximal somewhere between 100 and 150 m depth, although the exact depth and the predicted catch does vary between models (see Fig. 3, left panel). This amount of variation should be expected, as the inherent variability of the data is large for fisheries data. The relatively minor differences arise from different assumptions about the nature of the variability in the data. It is interesting to note that it is the DLNM that appears to differ the most, it is also the only model considered here that treats a zero observation as a qualitatively special value. Of course, it is impossible to know which of the four models best represents the truth as it is unknown.

The fitted mean-variance relationships for the different models are quite different although there are some general similarities); see Fig. 3 (right panel). The TGLM mean-variance relationship (a power relationship) gives fitted variances that are noticeably smaller than the other models'. The mean-variance relationship for the PGM-2 (a linear relationship—see (5) with constant $\mathrm{E}\left(w_{ij}\right)$) gives much larger variances. Both the PGM-1 and the DLNM show a relationship that varies with depth; see (1) and (5). The extra depth dependency is the cause of the loops in the relationship—the variance increases with expectation and depth but then decreases with expectation at a different rate due to different depths.

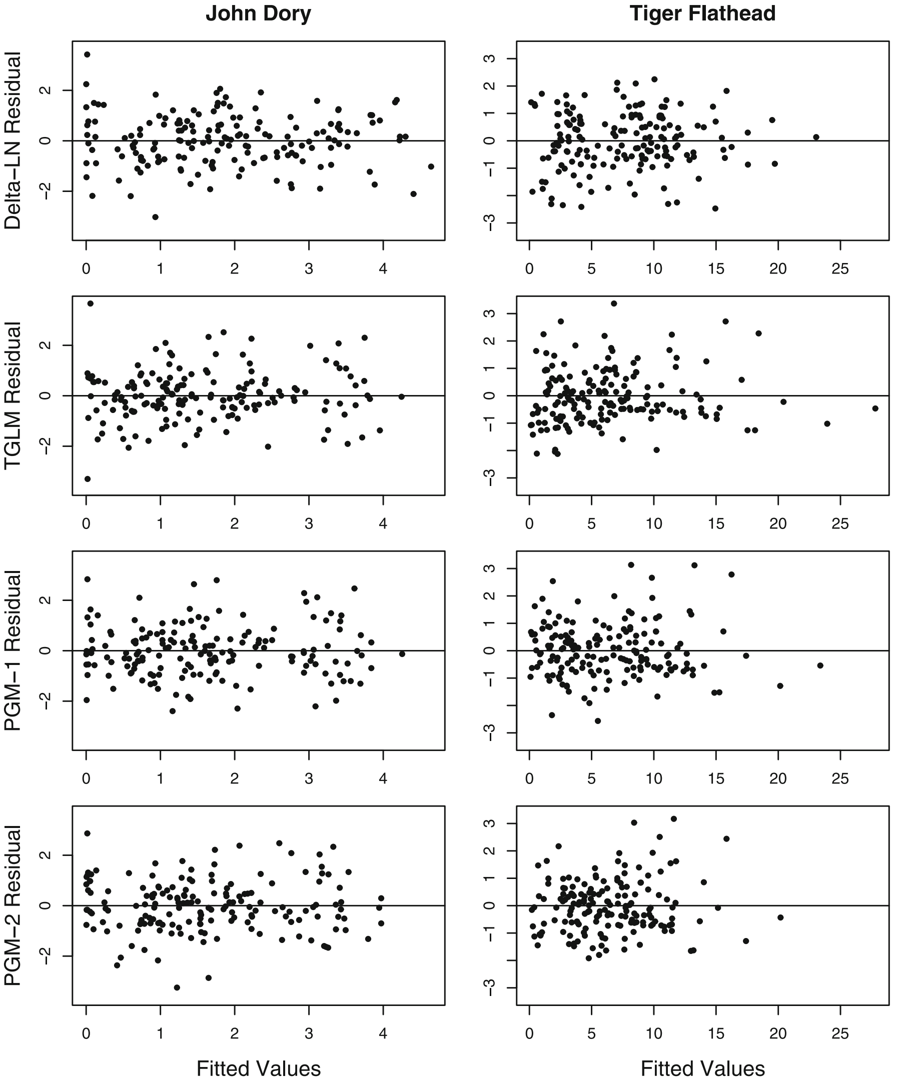

Fig. 2 Randomised quantile residuals for john dory (*left column*) and jackass morwong (*right column*). The four models are: delta log-normal (*top row*), Tweedie GLM (*second row*), Poisson–Gamma with depth affecting the summands (*third row*), and Poisson–Gamma without depth affecting the summands (*last row*)

4.2 The size of schools and individual fish

There were a number of species (5 out of the 15 that we analysed) where the depth dependence in the gamma part of the PGM appeared to advantageous. For these species, it is tempting to conclude that individual fish, and/or school size, vary with depth. The PGM alone can not distinguish between these three inferences. However, for the SEF we also have fish counts per trawl, which can be used to aid separation of the inferences. We now analyse these data assuming that a Poisson sums of gammas

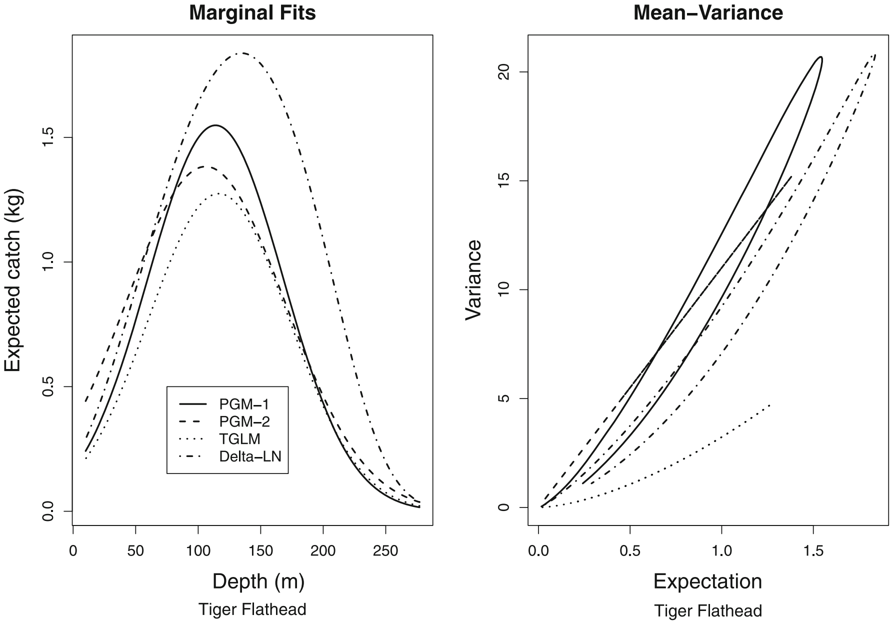

Fig. 3 Model summaries for tiger flathead from all four models. The along-coast variable is held constant at its mean and effort is held constant at one hectare of swept area. *Left panel* gives the expectations as a function of depth. *Right panel* gives the fitted mean-variance relationship

model is correct *and* also assuming that the summands represent individual fish. This is performed by fitting two separate GLMs sequentially.

First, a Poisson GLM is fitted to the abundance data, with dependence on depth and along-coast, and an offset for log area swept. The parameters estimated from this GLM should match those estimated from the Poisson part of the PGM if the summands represent individual fish.

Conditional on the abundances, if the summands are individual fish, the total weight for the ith sample will be distributed as $y_i|N_i \sim \text{gamma}(N_i\alpha, \beta_i)$. The weights of the individual fish can then be modelled as

$$\log\left(\mathrm{E}\left(y_i|N_i\right)\right) = \log\left(\alpha\beta_i\right) = \log N_i + \boldsymbol{x}_i^\top \boldsymbol{\tau},$$

which is a gamma GLM with $\log N_i$ as an offset. The samples that did not catch the species under consideration are not included in the GLM for weight; they have constant expectation, have no variance, and do not contribute to variation in the likelihood. Following the model for abundance, the estimated parameters from this model should be directly comparable to those obtained from the gamma summand part of the PGM.

The estimated parameters from the fits to the tiger flathead and jackass morwong weight *and* abundance data are given in Table 2. There is disagreement between the estimates for the different models. The intercept estimates suggest that the Poisson–Gamma model identifies that there are few schools and that those schools are larger. This suggests that the Poisson–Gamma model may be identifying variation in school

Table 2 Parameter estimates for (1) the Poisson–Gamma model fitted to the total weight data along, (2) the Poisson model for the abundance data alone, and (3) the total weight data conditional on abundance. There are two species, jackass morwong and tiger flathead

Model part	Covariate	Term		Tiger Flathead			Jackass Morwong	
			PGM[a]	Poisson[b]	Gamma[c]	PGM[a]	Poisson[b]	Gamma[c]
Poisson	Depth	Intercept	−1.37 (0.01)	0.82 (<0.01)	–	−2.43 (0.02)	0.93 (<0.01)	–
		Linear	4.06 (2.78)	15.33 (0.43)	–	−8.15 (3.43)	−6.39 (0.08)	–
		Quadratic	−6.21 (3.30)	−14.29 (0.32)	–	−4.13 (3.03)	−1.99 (0.06)	–
	Along-coast	Linear	0.04 (0.99)	−3.54 (0.04)	–	3.16 (2.39)	3.60 (0.11)	–
		Quadratic	2.27 (0.96)	3.41 (0.04)	–	−1.80 (2.33)	−6.22 (0.11)	–
		Cubic	0.77 (1.09)	1.78 (0.05)	–	−1.39 (2.61)	4.66 (0.09)	–
Gamma	Depth	Intercept	1.19 (0.02)	–	−0.98 (<0.01)	1.70 (0.04)	–	−0.70 (0.01)
		Linear	−0.89 (3.05)	–	−4.49 (0.26)	−11.67 (8.62)	–	−2.73 (0.40)
		Quadratic	−4.64 (3.50)	–	−1.01 (0.26)	−3.33 (6.26)	–	−0.19 (0.40)

[a] Poisson–Gamma model fitted to the total weight data only [b] Poisson GLM fitted to abundance data only [c] Gamma model for total weight conditional on abundance

size, or school size and fish size, rather than individuals. The estimates of the covariate parameters also disagree between the models. This could imply that the size and number of schools having different relationships to those for the size and number of individual fish. However, this claim is unprovable without information about the size and number of schools. The standard errors from the Poisson–Gamma model are much larger than those from the other models. This is expected as the Poisson and gamma models incorporate extra information through the abundance data.

5 Summary and discussion

In this manuscript we have presented a model to analyse data that is continuous and includes observations that are identically zero. Such data arise from many sources, including our motivation (fisheries data). Our model is based on the sum of gamma variates, of which there are a Poisson number. This simple and intuitive description of the data generating process is appealing as it provides interpretation that other models do not. However, that is not to say that an appealing interpretation is necessarily correct; if the interpretation is critically important, then it behooves the modeller to inspect other data, such as abundances (see Sect. 4.2).

The Poisson–Gamma model is closely related to the Tweedie GLM (Smyth 1996) but it has extra flexibility due to the ability to specify a non-constant mean-variance relationship. It also has certain similarities with a double GLM (DGLM; see Smyth and Verbyla 1999), where the mean and dispersion of a GLM are unconnected link-linear functions of covariates. We think that the PGM offers a more natural method for continuous non-negative data as there is a sensible, albeit simplistic, generating mechanism. However, the PGM is likely to offer a less flexible mean-variance structure than a DGLM. This is because the parameters in the Poisson–Gamma model do not solely control mean or variance as the parameters do in a double GLM; rather, they control both.

Any of the models considered in this manuscript could behave poorly when there is a high proportion of zeros. This is because the zero observations lack variation amongst themselves. Consideration of the Poisson–Gamma model highlights the reason—when there are many zeros the amount of information in the non-zero data is small. This is the only information used to estimate the gamma part of the model, but it is also has to be shared with the Poisson part of the model.

The proposed model appears to perform well in a fisheries setting; see Sect. 4.1. Residual plots showed that the Poisson–Gamma model fitted the data for 15 fish species where the fit was as good as the Tweedie GLM and the delta log-normal model, which are both commonly used. This raises the question about when the Poisson–Gamma model should be used. The answer is when at least one of the following is required: (1) the mean-variance relationship varies with covariates, (2) the zero class should not be treated as a special case, and (3) an overall multiplicative model is required. The first situation distinguishes it from a TGLM and the second two distinguishes it from a delta log-normal model. Another distinction between the PGM and the TGLM is the ability to add an offset that affects the mean and variance in a linear manner. This is not possible with the TGLM. This attribute is appealing for samples in space where an area sampled offset is likely to affect the number of summands encountered but unlikely to affect the summands size.

Acknowledgments This work was partly funded by the Fisheries Research and Development Corporation, project 2008/002. We would like to thank Nic Bax, Piers Dunstan, David Peel, and an anonymous referee for advice on data and/or presentation of the manuscript.

A Calculating the Poisson–Gamma and Tweedie densities and their derivatives (web-Appendix)

A.1 Log densities

The main challenge for evaluating the probability density function (pdf) for the Poisson–Gamma model and for the Tweedie model is the evaluation of the generalised Bessel function (6). The methods outlined here follow those described in Dunn and Smyth (2005). Some minor differences arise due to the Poisson–Gamma parameterisation rather than the Tweedie parameterisation. In this Appendix we derive the expressions and computational strategies for calculating the densities. We omit the more tedious algebra, of which there is plenty. For observation y and parameters (λ, β, α) the function under consideration is

$$\begin{aligned}\log\left(W(y,\lambda,\alpha,\beta)\right) &= \log\left(\sum_{j=1}^{\infty}\frac{\lambda^j\left(\frac{y}{\beta}\right)^{j\alpha}}{j!\Gamma(j\alpha)}\right)\\ &= \log\left(\sum_{j=1}^{\infty} w_j\right) &(7)\\ &= \log\left(w_*\right) + \sum_{j=1}^{\infty}\exp\left[\log\left(w_j\right) - \log\left(w_*\right)\right], &(8)\end{aligned}$$

where w_* is the maximum of the set of w_j. The form (7) is preferred to (8) for computation as it reduces the risk and effect of numerical underflow. Each of the w_j terms in (7) can be calculated using readily-available routines for calculating the log-gamma function. A common but not always accurate approximation is the one based on Stirling's approximation (see Abramowitz and Stegun 1964), which gives necessary terms to be

$$\begin{aligned}\log w_j &\approx j\left[\log\lambda + \alpha\log y - \alpha\log\beta - \alpha\log\alpha + 1 - (\log(j+1) + \alpha\log j)\right] \\ &\quad -\frac{1}{2}\left[\log(j+1) - \log j\right] + j\alpha + \frac{1}{2}\log\alpha - \log 2\pi + 1 \\ &= j\left[z_1 - (\log(j+1) + \alpha\log j)\right] - \frac{1}{2}\left[\log(j+1) - \log j\right] + j\alpha + z_2, \text{ say.} \quad (9)\end{aligned}$$

The infinite series 7 is approximated using the idea from Dunn and Smyth (2005) where the maximum term (w_*) is located by simple calculus. Treating j as continuous over $[1, \infty)$ the first derivative is

$$\begin{aligned}\frac{\partial\log w_j}{\partial j} &\approx z_1 - \log(j+1) - \alpha\log j - \frac{2j+1}{2(j+1)} + \frac{1}{2j} \\ &\approx z_1 - (1+\alpha)\log j - 1.\end{aligned}$$

The second approximation becomes more accurate as j becomes large. The second derivative is (approximately) negative for all j implying that there is a single maxima. This maxima is found at

$$j_m \approx \exp\frac{z_1 - 1}{1+\alpha}.$$

However, this expression gives a real number not an integer. We choose j_* to be that associated with the larger of $\log w_{\lfloor j_m \rfloor}$ and $\log w_{\lceil j_m \rceil}$.

Finally, the approximation is calculated by stepping out from j_*, in both directions, until the w_j are sufficiently small with respect to w_{j_*}. We implement this to be the set of integers $\{j\}$ such that $\log w_j - \log w_* > -37$, which gives accuracy to about machine error.

A.2 First derivatives

The first derivatives of the log-likelihood contributions (6) are easily obtained, except for the derivative of $\log W$, with respect to all parameters. We change parameterisation to (λ, μ, α) where $\mu = \alpha\beta$ and is the expectation of the individual gamma variables in the Poisson–Gamma formulation; see Sect. 2.3. Each term in the summation for the generalised Bessel function is then

$$w_j = \frac{\lambda^j\left(\frac{y\alpha}{\mu}\right)^{j\alpha}}{j!\Gamma(j\alpha)}.$$

The change of parameters is performed as the μ parameter, not β, that is directly modelled in the Poisson–Gamma model. Using λ as an example, the derivative is

$$\frac{\partial\log w}{\partial\lambda} = \frac{1}{w}\frac{\partial w}{\partial\lambda} = \frac{1}{w}\sum_{j=1}^{\infty}\frac{\partial w_j}{\partial\lambda}$$

and the other derivatives follow similarly. The series sum, W, is obtained using the methods outlined in Appendix A.1. The derivatives of the series terms are given by

$$\begin{aligned}\frac{\partial w_j}{\partial \lambda} &= \frac{j}{\lambda} w_j\\ \frac{\partial w_j}{\partial \mu} &= -\frac{j\alpha}{\mu} w_j\\ \frac{\partial w_j}{\partial \alpha} &= j w_j \left[1 + \log\left(\frac{y\alpha}{\mu}\right) - \Psi(j\alpha)\right]\end{aligned}$$

where $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ is the digamma function. Note that the derivative with respect to λ is always positive, the derivative with respect to μ is always negative and the derivative with respect to α may be either. Following Appendix A.1 and Dunn and Smyth (2005) we treat j as continuous and differentiate the log of the derivatives of w_j. That is

$$\frac{\partial}{\partial j} \log\left(\frac{\partial w_j}{\partial \lambda}\right) = \tfrac{\partial}{\partial j} \log\left(-\tfrac{\partial w_j}{\partial \mu}\right) = \tfrac{1}{j} + \tfrac{\partial \log w_j}{\partial j}$$

and we ignore the derivative of $\frac{\partial w_j}{\partial \alpha}$ for now. Note that the maximum of both the derivatives with respect to λ and μ will have an optimum near j_*, the maximum of the w_j series, especially if j is large. Following the (approximate) second derivative for w_j with respect to j, it is seen that $\frac{\partial w_j}{\partial \lambda}$ and $\frac{\partial w_j}{\partial \mu}$ will be negative. Hence, computation can follow that of Appendix A.1.

Calculation of $\frac{\partial w_j}{\partial \alpha}$ is performed in a similar manner, but is performed on the absolute values of the individual summands that are subsequently sign corrected. This is done, once again, to reduce numerical underflow. The summation is performed as a finite series approximation to

$$\begin{aligned}S_\alpha &= \sum_{j=1}^{\infty} \text{sgn}\left(\frac{\partial w_j}{\partial \alpha}\right) \text{sgn}\left(\eta_*\right) \exp\left(\log\left|\frac{\partial w_j}{\partial \alpha}\right| - \log|\eta_*|\right), \text{ and}\\ \frac{\partial W}{\partial \alpha} &= \text{sgn}\left(\eta_*\right) \text{sgn}\left(W_\alpha^*\right) \exp\left(\log|\eta_*| + \log|S_\alpha|\right),\end{aligned}$$

where $\text{sgn}\left(\cdot\right)$ is the sign function and $\eta_* = \max\left(\left|\frac{\partial w_j}{\partial \alpha}\right|\right)$.

The justification for the simple summing procedure for $\frac{\partial W}{\partial \alpha}$ is slightly more involved and requires some proof that the series is converging (to zero) as j increases. First, note that the only term in the derivative that may not converge to zero is $j w_j \Psi(j\alpha)$ as all others are the ratio of a linearly-increasing term and a super-linearly-increasing term, which converges to zero. The digamma function is monotonically increasing on the positive real numbers. Further, it is increasing a a rate which is sub-linear (see recurrence relation in Abramowitz and Stegun 1964, p. 258). The ratio of these terms

Table 3 Summary of parameter estimates for the simulation study for Tweedie GLM ($n = 10,000$)

Covariate	Term	Value	Estimate (±SE[a])	SD[b]	ASD[c] (±SE)
Depth	Intercept	−0.879	−0.975 (0.002)	0.216	0.204 (<0.001)
	Linear	−22.121	−22.288 (0.028)	2.753	2.648 (0.002)
	Quadratic	−7.648	−8.090 (0.025)	2.452	2.369 (0.002)
Along-coast	Linear	5.749	5.874 (0.026)	2.584	2.409 (0.002)
	Quadratic	−4.116	−4.701 (0.026)	2.590	2.412 (0.003)
	Cubic	4.660	4.681 (0.025)	2.547	2.422 (0.003)
Dispersions	ϕ	7.461	7.238 (0.008)	0.778	0.777 (0.001)
	p	1.656	1.646 (0)	0.031	0.030 (<0.001)

[a] Empirical standard error of the mean of the 10,000 parameter estimates. [b] Empirical standard deviation of the 10,000 parameter estimates. [c] Mean asymptotic standard deviation of the parameter estimates (standard error)

is thus converging to zero, irrespective of sign. Hence, all terms converge to zero, as does their sum.

A.2.1 Derivatives for poisson–Gamma model and Tweedie GLM parameters

Appendix A.2 gave methods for evaluating the derivative of the Tweedie density with respect to the parameters of the Poisson–Gamma model. Under the Poisson–Gamma model two of these parameters (λ and μ) are allowed to vary with covariates. The derivatives with respect to the parameters associated with the covariates is found by a simple application of the chain rule. For example, the derivative of the density with respect to the parameters for the Poisson mean ($\boldsymbol{\tau}$) is

$$\begin{aligned}\frac{\partial \log f(y)}{\partial \boldsymbol{\tau}} &= \frac{\partial \log f(y)}{\partial \lambda}\frac{\partial \lambda}{\partial \eta}\frac{\partial \eta}{\partial \boldsymbol{\tau}} \\ &= \frac{\partial \log f(y)}{\partial \lambda}\lambda \boldsymbol{x}^{\top}\end{aligned}$$

where η is the linear predictor, and $\boldsymbol{x}$ is the column vector of covariates. The second line follows due to the log-link function.

The derivatives for the Tweedie GLM are found in a similar manner but require an extra derivative to account for the re-parameterisation. That is for the Tweedie GLM's location parameters ($\boldsymbol{\tau}$)

$$\frac{\partial \log f(y)}{\partial \boldsymbol{\tau}} = \frac{\partial \log f(y)}{\partial(\lambda,\mu,\alpha)}\frac{\partial(\lambda,\mu,\alpha)}{\partial \mathrm{E}(y)}\frac{\partial \mathrm{E}(y)}{\partial \eta}\frac{\partial \eta}{\partial \boldsymbol{\tau}}.$$

The entries of the second term are obtained by differentiating the relationships in (3). Similar expressions for the Tweedie GLM's dispersion (ϕ) and power parameter (p) are readily obtained in an analogous manner.

B Simulation results for Tweedie GLM (web-Appendix)

The results of the simulation study that investigates the estimability of the Tweedie GLM's parameters are now given; see Table 3. The simulation experiment had a very similar design to that for the Poisson–Gamma model, except (of course) the parameter values used for simulation were taken from a Tweedie GLM, not a Poisson–Gamma model.

There appears to be some mild bias in most parameters. The simulation study shows that the bias is statistically important but, given the size of the dominant parameter for the polynomial, this level of bias is practically inconsequential. The estimates' standard errors appear to be slightly underestimated from the asymptotic values; see Table 3, and the covariances are similarly deflated.

References

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions with formulas, graphs, and mathematical tables, ninth Dover printing, tenth GPO printing edn. Dover, New York

Aitchison J (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. J Am Stat Assoc 50(271):901–908

Ancelet S, Etienne M, Benot H, Parent E (2010) Modelling spatial zero-inflated continuous data with an exponentially compound poisson process. Environ Ecol Stat 17(3):347–376

Bax N, Williams A (eds) (1999) Habitat and fisheries production in the South East fishery. Final report to FRDC Project 94/040, CSIRO Marine Research, Hobart

Brynjarsdottir J, Stefansson G (2004) Analysis of cod catch data from icelandic groundfish surveys using generalized linear models. Fish Res 70(2–3):195–208

Candy SG (2004) Modelling catch and effort data using generalised linear models, the tweedie distribution, random vessel effects and random stratum-by-year effects. CCAMLR Sci 11:59–80

Dick E (2004) Beyond 'lognormal versus gamma': discrimination among error distributions for generalized linear models. Fish Res 70:351–366

Dunn P, Smyth G (1996) Randomized quantile residuals. J Comput Graph Stat 5(3):236–244

Dunn P, Smyth G (2005) Series evaluation of tweedie exponential dispersion model densities. Stat Comput 15:267–280

Heilbron D (1994) Zero-altered and other regression models for count data with added zeros. Biometr J 36:531–547

Jørgenson B (1997) The theory of dispersion models. Chapman and Hall, London

Kendal W (2004) Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. Ecol Complex 1:193–209

Kendal W, Jørgensen B (2011) Taylors power law and fluctuation scaling explained by a central-limit-like convergence. Phys Rev E 83

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

Maunder MN, Punt AE (2004) Standardizing catch and effort data: a review of recent approaches. Fish Res 70(2–3):141–159

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton

Mullahy J (1986) Specification and testing of some modified count data models. J Econ 33:337–350

Muralidharan K, Kale B (2002) Modified gamma distribution with singularity at zero. Commun Stat Simul Comput 31(1):143–158

Muralidharan K, Lathika P (2007) Statistical modelling of rainfall data using modified Weibull distribution. MAUSAM Indian J Meteorol Geophys Hydrol 56(4):765–770

Nash S, Sofer A (1996) Linear and nonlinear programming. McGraw-Hill, Singapore

Ortiz M, Arocha F (2004) Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: Billfish species in the venezuelan tuna longline fishery. Fish Res 70(2–3):275–297

Peel D, Bravington M, Kelly N, Wood S, Knuckey I (2012) A model-based approach to designing a fishery independent survey. J Agric Biol Environ Stat (in press)

Punt AE, Walker TI, Taylor BL, Pribac F (2000) Standardization of catch and effort data in a spatially-structured shark fishery. Fish Res 45(2):129–145

Ridout M, Demétrio C, Hinde J (1998) Models for count data with many zeros. In: Invited paper presented at the nineteenth international biometric conference, Cape Town, pp 179–190

Shono H (2008) Application of the tweedie distribution to zero-catch data in cpue analysis. Fish Res 93:154–162

Smyth G (1996) Regression modelling of quantity data with exact zeros. In: Proceedings of the second Australia-Japan workshop on stochastic models in engineering. Technology and Management, Technology Management Centre, University of Queensland, pp 572–580

Smyth G, Verbyla A (1999) Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics 10:695–709

Stefansson G (1996) Analysis of groundfish survey abundance data: combining the glm and delta approaches. Ices J Marine Sci 53(3):577–588

Tascheria R, Saavedra-Nievasb J, Roa-Ureta R (2010) Statistical models to standardize catch rates in the multi-species trawl fishery for patagonian grenadier (Macruronus magellanicus) off southern chile. Fish Sci 105:200–214

Taylor L (1961) Aggregation, variance and the mean. Nature 189:732–735

Thompson R, Baker R (1981) Composite link functions in generalized linear models. J R Stat Soc Ser C 30(2):125–131

Warton D (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16:275–289

Author Biographies

Scott D. Foster is a statistician with CSIRO's Division of Mathematics, Informatics and Statistics. His interests lay in statistical methods for solving problems arising from marine ecology. This encompasses problems from fisheries, community ecology and biodiversity research.

Mark V. Bravington is a senior statistician with CSIRO's Division of Mathematics, Informatics and Statistics. His research is driven by problems arising from marine ecology. This guides him to problems of animal abundance, survey design, animal movement, amongst others. He has a strong interest in statistical computing.