Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 1159–1166 doi: 10.1111/2041-210X.12122

Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume

Jean-Baptiste Lecomte[1,2]*, Hugues P. Benoît[3], Sophie Ancelet[4], Marie-Pierre Etienne[1,2], Liliane Bel[1,2] and Eric Parent[1,2]

[1]INRA, UMR 518 Math. Info. Appli., Paris F-75005 France; [2]AgroParisTech, UMR 518 Math. Info. Appli., Paris F-75005, France; [3]Gulf Fisheries Centre, Fisheries and Oceans Canada, Moncton, NB E1C 9B6,Canada; and [4]Institut de Radioprotection et de Sûreté Nucléare, Laboratoire d'épidémiologie, Fontenay-aux-Roses, France

Summary

1. Ecological data such as biomasses often present a high proportion of zeros with possible skewed positive values. The Delta-Gamma (DG) approach, which models separately the presence–absence and the positive biomass, is commonly used in ecology. A less commonly known alternative is the compound Poisson-gamma (CPG) approach, which essentially mimics the process of capturing clusters of biomass during a sampling event.

2. Regardless of the approach, the effort involved in obtaining a sample (henceforth called the sampling volume, but could also include swept areas, sampling durations, etc.), which can potentially be quite variable between samples, needs to be taken into account when modelling the resulting sample biomass. This is achieved empirically for the DG approach (using a generalized linear model with sampling volume as a covariate), and theoretically for the CPG approach (by scaling a parameter of the model). In this study, the consequences of this disparity between approaches were explored first using theoretical arguments, then using simulations and finally by applying the approaches to catch data from a commercial groundfish trawl fishery.

3. The simulation study results point out that the DG approach can lead to poor estimates when far from standard idealized sampling assumptions. On the contrary, the CPG approach is much more robust to variable sampling conditions, confirming theoretical predictions. These results were confirmed by the case study for which model performances were weaker for the DG.

4. Given the results, care must be taken when choosing an approach for dealing with zero-inflated continuous data. The DG approach, which is easily implemented using standard statistical softwares, works well when the sampling volume variability is small. However, better results were obtained with the CPG model when dealing with variable sampling volumes.

Key-words: commercial fishery catches, compound Poisson, estimation of biomass, sampling variability, two-part model

Introduction

Ecological data for species population densities are often characterized by a large proportion of zero values accompanied by a skewed distribution of remaining values, including occasional extremes (Pennington 1996; Martin *et al.* 2005). Ignoring these features could lead to incorrect estimates of quantities of interest (e.g. mean biomass, probability of presence) and their associated uncertainty, and possibly to incorrect conclusions (Martin *et al.* 2005). Zero values in species population densities can originate from two general sources, with consequences for the appropriate analytical approach used to make inferences [see review in Martin *et al.* (2005)]. *True zeros* can occur as a direct result of the effect under study (e.g. suitability of a given habitat) or as a stochastic result of sampling from areas of low density. On the other hand, *false zeros* can occur as a result of detection limits or observer effects. Our interest here lies in *true zeros*.

Standard continuous probability distributions such as the normal, gamma or log-normal are often inappropriate for the analysis of zero-inflated biomass data, even with ad hoc assumptions such as the addition of constants to create a mass at zero. A better approach is to use so-called two parts, hurdle or Delta models, which assume that zero and nonzero data arise, respectively, from separate processes (Stefansson 1996; Punt *et al.* 2000; Ortiz & Arocha 2004; Maunder & Punt 2004). This method does not require the addition of a constant, which can introduce a bias in the data. This model is also very flexible as covariates can be added in the zero and nonzero parts of the model using conventional generalized linear modelling techniques. However, the break between zero and

*Correspondence author. E-mail: jean-baptiste.lecomte@agroparistech.fr

© 2013 The Authors. Methods in Ecology and Evolution © 2013 British Ecological Society

nonzero values presents a particularly unnatural discontinuity in density data, where many zeros are actually stochastic clues of a strong gradient of decreasing biomass quantities. A second approach is the use of a positive distribution that simultaneously incorporates zeros and positive quantities. Jorgensen (1987) proposed the exponential dispersion model, with a power variance function. This model, also known as the Tweedie distribution, handles zero-inflated data without treating the zero and nonzero values separately. The Tweedie model and its variants have been applied to fisheries data (Candy 2004; Shono 2008; Foster & Bravington 2012; Lecomte *et al.* 2013). In this article, we rely on a gamma marked compound Poisson, named compound Poisson-gamma model (CPG), a member of the Tweedie family. Foster & Bravington (2012) extended it to be more flexible when covariates can affect parameters. They showed that the CPG mean–variance relationship is not necessarily constant, conversely to the Tweedie distribution (Foster & Bravington 2012). Parsimonious variant of this distribution, using exponential rather than gamma variables, has also been used [e.g. Ancelet *et al.* (2010)].

In many studies, the effort involved in obtaining a sample (henceforth called the sampling volume, but could also include swept areas, sampling durations, etc) can vary among sampling events. These differences in the sampling volume have to be accounted for in the analysis. Variable sampling volume is accounted for directly in the modelling for the CPG approach by scaling a parameter, whereas recourse to a generalized linear model to take into account the sampling volume as a covariate or an offset is required for the delta-gamma (DG) approach (Maunder & Punt 2004). Such different approaches to dealing with variable sampling volumes are likely to affect estimation reliability for quantities of interest (e.g. mean quantity, probability of presence).

This study evaluates the relative robustness of the DG and CPG approaches for estimating biomasses and presence probabilities under variable sampling volumes conditions in three ways. Firstly, the form and analytical properties of the two models are presented and contrasted from a theoretical perspective. Secondly, simulations were used to evaluate the robustness of the proposed models and compare their fitting abilities with variable volumes with different variances. Thirdly, the two approaches are applied to catch data from a commercial groundfish trawl fishery. Theory and analyses of simulated and observed data have all indicated that the CPG approach outperforms the DG approach under variable sampling volumes.

Materials and methods

THE DELTA-GAMMA MODEL

The delta modelling approach is based on the specification of two submodels to represent the biomass (Stefansson 1996). Let X be a binary variable that equals to 1 if the species of interest is present and 0 otherwise.

$$X \sim Bernoulli(\pi) \qquad \text{eqn 1}$$

π being the probability of the presence of the species. Conditionally, let Y be a positive sampled quantity of interest (e.g. species density or biomass) after a sampling event:

$$\begin{aligned} Y|X=1 &\sim \Gamma(\alpha,\beta) \\ Y|X=0 &\sim \delta_0 \end{aligned} \qquad \text{eqn 2}$$

with shape and rate parameters, (α,β) and δ_0 the Dirac distribution at zero. This yields the DG model, DG(π,α,β), with other distributional assumptions for strictly positive quantities yielding other models in the delta family, such as the delta log-normal. The expected value for the biomass under the DG model is $\mathbb{E}(Y) = \frac{\alpha}{\beta}\pi$, and the other main derived quantities (e.g. variance of the biomass, probability of presence) are summarized in Table 1.

Table 1. Quantities of interest (probability of presence, expected positive biomass, expected biomass and variance of biomass) for the DG and the CPG model under a standard sampling volume

	DG	CPG
Probability of presence	π	$1-e^{-\lambda}$
Expected positive biomass	$\frac{\alpha}{\beta}$	$\frac{a\lambda}{b(1-e^{-\lambda})}$
Expected biomass	$\frac{\alpha}{\beta}\pi$	$\lambda\frac{a}{b}$
Variance of the biomass	$\pi\frac{\alpha}{\beta^2}(1+\alpha(1-\pi))$	$\frac{a\lambda}{b}\frac{a+1}{b}$

A useful model property in statistical ecology is additivity with regards to the sampling volumes, in which the sum of two independent sampling events follows the same distribution type as each sampling event. For example, sampling during two hours follows the same distribution as two samplings process of one hour. It allows gathering data with different sampling volumes in the same model, as their sum is obtained according to a distribution in the same family. Unfortunately, the DG model is not additively coherent as pointed out by Stefansson (1996). As a consequence, it is not clear how the DG parameters vary in time or space when sampling volumes vary among sampling events. In practice, a simple way to deal with a non-constant sampling volume is to perform a pre-standardization of the data. The biomass collected is divided by the sampling volume. The problem with this method is that it only standardizes the positive data and leaves the presence–absence part unscaled, ignoring the fact that, as sampling volume increases, the probability of observing zero biomass should decrease when the species is present. A more relevant solution uses generalized linear modelling (e.g. Zuur *et al.* 2009) with the sampling volume as a covariate in each part of the DG model. The probability of presence is usually modelled with a logistic regression:

$$\log\left(\frac{\pi}{1-\pi}\right) = \zeta + \phi V \qquad \text{eqn 3}$$

where V is the sampling volume. A log-linear function is used to add the sampling volume in the expected positive biomass given the species is present :

$$\mathbb{E}(Y|X=1) = \frac{\alpha}{\beta} = \exp(\gamma + \delta V) \qquad \text{eqn 4}$$

A criticism of the delta approach is the separate modelling of the presence–absence and the strictly positive quantities. Consequently, gradients of biomass including low or null densities of the species are modelled disjointly in practice, which is a rather unnatural representation of the phenomenon being modelled.

THE COMPOUND POISSON-GAMMA MODEL

Conceptually, the CPG mimics the process involved when sampling most living organisms in nature for which the observed variable of interest is a continuous variable, such as the total

biomass captured during a sampling event (Foster & Bravington 2012; Lecomte *et al.* 2013). Simply put, the model assumes that a Poisson distributed number N of aggregations (i.e. patches or lumps) of organisms are collected, each patch containing a mass M_p modelled using a gamma distribution. It should be noted that an aggregation could contain only one organism (Foster & Bravington 2012). The sum of the individual masses of captured aggregations yields the total observed biomass Y:

$$Y = \begin{cases} \sum_{p=1}^{N} M_p & \text{if} \quad N > 0 \\ 0 & \text{if} \quad N = 0 \end{cases} \qquad \text{eqn 5}$$

The CPG is characterized by three parameters: λ the Poisson intensity, a and b the shape and rate gamma parameters:

$$Y \sim \mathrm{CPG}(\lambda, a, b) \qquad \text{eqn 6}$$

The main derived quantities for the CPG model are summarized in Table 1.

Due to additivity properties (Jorgensen 1987), the sampling volume V may be straightforwardly incorporated in a CPG model by scaling the Poisson intensity parameter:

$$Y \sim \mathrm{CPG}(\lambda V, a, b) \qquad \text{eqn 7}$$

The CPG approach jointly models the probability of presence and the nonzero sampled quantity. This capacity allows one to model a gradient of decreasing biomass in the distribution of the targeted species due to low density of organisms or low detectability.

There is no disjoint treatment of null and positive values as in the DG model. Foster & Bravington (2012) note that when no covariates are included in either the Poisson or gamma latent components, the CPG model belongs to the Tweedie family, and, in addition, a reviewer has noted that this is still the case when the set of covariates is identical in each of the Poisson and gamma components.

A SIMULATION STUDY TO COMPARE THE IMPACT OF VARIABLE SAMPLING VOLUME

The abilities of the DG and CPG models to reliably estimate quantities of interest when sampling volume is variable were compared using simulations. The trawls are divided into small fractions, or microvolumes, that could conceptually be although of as the sweeping of one unit of area by the trawl. Each microvolume contains a small amount of biomass produced according to a DG process. The observed sampled biomass is the sum of the biomass collected over those small microvolumes. Because the DG does not possess additivity, a biomass amount summed over all microvolumes constituting a complete trawl haul does not conform with either the DG or CPG model. The simulation proceeded as follows:

1. Biomass values were generated with a DG model of parameters α_{micro}, β_{micro} and π_{micro} corresponding to a sampled fraction V_{micro}. These biomasses are denoted as 'microbiomasses'.

2. The total collected biomass of a sample is the sum of N_{V} microbiomasses captured across all sampled microvolumes for that sample to result in a total volume V:

$$N_{\mathrm{V}} = \frac{V}{V_{\mathrm{micro}}}$$

3. The total volumes V are simulated according to a log-normal distribution:

$$V \sim \log N(0, \sigma^2)$$

with one of several variances σ^2 which varied between simulations: 0·1, 0·2, 0·3, 0·5, 0·7, 0·9, 1.00, corresponding, respectively, to a coefficient of variation of the sampling volumes C_{V}: 0·10, 0·20, 0·31, 0·53, 0·80, 1·12, 1·31 and a constant median of 1. For each σ^2 value, $n = 100$ data sets composed of 150 full samples were generated of the population of microvolumes.

Three quantities of interest can be expressed analytically as a function of the microvolume parameters. The expected biomass collected over N_{V} microvolumes, each producing a microbiomass from a DG distribution with parameters (α_{micro}, $\beta_{\mathrm{micro}}\pi_{\mathrm{micro}}$), equals:

$$Q = \mathbb{E}(Y) = \sum_{i=1}^{N_{\mathrm{V}}} \mathbb{E}(Y_{\mathrm{micro}_i}) = N_{\mathrm{V}}\pi_{\mathrm{micro}}\frac{\alpha_{\mathrm{micro}}}{\beta_{\mathrm{micro}}}.$$

The probability π of presence is obtained by noticing that:

$$\pi = P(Y > 0) = 1 - P(Y = 0) = 1 - \prod_{i=1}^{N_{\mathrm{V}}} P(Y_{\mathrm{micro}_i} = 0) = 1 - (1 - \pi_{micro})^{N_{\mathrm{V}}}. \qquad \text{eqn 8}$$

Finally, the strictly positive expected biomass is given by:

$$\mathrm{QP} = \mathbb{E}(Y|Y > 0) = \frac{\mathbb{E}(Y)}{P(Y > 0)} = N_V\pi_{\mathrm{micro}}\frac{\alpha_{micro}}{\beta_{\mathrm{micro}}(1 - (1 - \pi_{\mathrm{micro}})^{N_{\mathrm{V}}})}.$$

To simulate zero-inflated biomass data with a variation in the sampling volume, a very small microvolume $V_{\mathrm{micro}} = 0{\cdot}001$ and large numbers N_{V} were considered. According to equation 8, π_{micro} has to be chosen very small to simulate a realistic probability of presence. Three contrasting sets of the parameters (α_{micro}, β_{micro}, π_{micro}) were considered as follows: (200,2,0·001), (200,2,0·0005), (20,2,0·001). In those cases, if $N_{\mathrm{V}} = 1000$, the resulting sampled volume is $V = 1$. Thus, when $N_{\mathrm{V}} = 1000$, the mean biomass of a data set generated with parameters (200,2,0·001) was $Q = 100$, and the probability of presence was $\pi = 0{\cdot}63$. This data set presented a reasonable proportion of zeros associated with large positive biomasses that is often encountered in ecological surveys. Data sets generated with parameters (200,2,0·0005) were intended to investigate higher proportion of zeros ($Q = 50$ and $\pi = 0{\cdot}39$), whereas data sets simulated with parameters (20,2,0·001) were representative of situation with lower quantities of biomass ($Q = 10$ and $\pi = 0{\cdot}63$). Summing over a large number of microvolumes N_{V} allowed to simulate realistic continuous zero-inflated data with a variation in the sample volume. However, one may object that the previous large sum of microvolumes could unduly favour the additively consistent CPG model. That is why, a fourth set of parameters with a larger π_{micro} was chosen to test the robustness of the CPG model in situation far from the addition of a large number of microvolumes. In this case, a larger microvolume was chosen $V_{\mathrm{micro}} = 0{\cdot}3$, and a small number of microvolumes N_{V} were summed to ensure a realistic overall probability of presence. When $N_{\mathrm{V}} = 3$, data sets generated with this set of parameters (200,2,0·15) presented a mean biomass $Q = 45$ and a probability of presence $\pi = 0{\cdot}39$.

BAYESIAN INFERENCE

We choose to use Bayesian inference and computation, using Markov chain Monte Carlo methods. For both models, the Bayesian model specification requires prior distributions. Commonly, vague normal distributions, with mean zero and standard deviation 100, were chosen for

all regression parameters. For the positive parameters, weakly informative gamma prior distributions, *Gamma*(1,0.001), were chosen. The inference was carried out using OpenBUGS, the open version of WinBUGS (Ntzoufras 2011). For each model, three chains were run for 60,000 iterations, with a burn-in period of 30 000 iterations. A thinning of 100 iterations was performed to avoid autocorrelations in each chain. Convergence was assessed using the Gelman–Rubin convergence test. Maximum likelihood estimation of both models can be found in Foster & Bravington (2012).

MODEL EVALUATION

The effect of variable sampling volume was examined for three quantities of interest θ, namely mean biomass, mean strictly positive biomass and probability of absence. The results obtained for the two models were explored using four performances indices for each quantity. The first was the root mean squared error, which accounts for the common trade-off between variance and bias of the posterior mean of the quantity for the *ith* data set, $\widehat{\theta}_i$. It is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_{\mathrm{true}} - \widehat{\theta}_i)^2} \qquad \text{eqn 9}$$

where θ_{true} is the 'true' value of θ used in the simulations. The second was the estimated average coefficient of variation computed for each unknown quantity of interest, which highlights the relative estimated dispersion, and is defined as:

$$\hat{C}_{\mathrm{V}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\sigma}_{\theta_i}}{\widehat{\theta}_i} \qquad \text{eqn 10}$$

where $\widehat{\sigma}_{\theta_i}$is the posterior standard deviation of θ_i related to data set *i*. The third is the recovery ratio, $R_{90\%}$, (sometimes called the confidence coefficient), which is obtained by counting over the 100 data sets, how many times the true value falls within the 90% credible interval. It highlights the fitting capacity of the model. Finally, the average posterior median, $\bar{\widehat{\theta}}$, over the n=100 replicated data sets was computed as an estimator of the three quantities of interest.

CASE STUDY: COMMERCIAL FISHERY GROUNDFISH DATA

The consequences for how the CPG and DG approaches deal with variable sampling volumes were explored by applying the methods to commercial fishery catches, which are known to present variable sampling volumes between sampled sites and a high proportion of zeros. This case study is particularly pertinent because the synthesis of commercial fishery catches is routinely used to assess relative stock abundance in fisheries worldwide (e.g. Maunder & Punt 2004).

The data consisted of bottom trawl catches for two years, 2006 and 2009, from a commercial fishery that covered the continental shelf off the west coast of Canada. The two years of data were chosen because they presented a contrast in annual dispersions of the sampling duration. The mean duration of a sampling event for both years was 120 minutes, and all sampling volumes were scaled accordingly so that one unit of sampling effort corresponds to two hours of towing. Histograms of the sampling duration after rescaling by the mean are provided in Fig.1. The variation observed in these fisheries is commensurate with variation observed in other fisheries elsewhere (e.g. Fig.2). Such scaling by the mean led to the following contrasted variance between the selected years:

- 2006, with empirical variance $\hat{\sigma}^2_{V_{2006}} = 0{\cdot}31$ and empirical coefficient of variation $C_{\mathrm{V}} = 0{\cdot}56$,
- year 2009: with empirical variance $\hat{\sigma}^2_{V_{2009}} = 0{\cdot}14$ and empirical coefficient of variation $C_{\mathrm{V}} = 0{\cdot}37$.

The data for two species exhibited differences in mean sampled density between the dover sole (*Microstomus pacificus*, $Q_{\mathrm{sole}_{2009}} = 31$ in kg per tow) and the Pacific Ocean perch (*Sebastes alutus*, $Q_{\mathrm{perch}_{2009}} = 267$ in kg per tow). Both models were applied to data from each species and year separately. Depth (in metres) was added to both models as a covariate to account for its well-known effect on catch rates. Depth, which ranged from 50 to 500 *m*, was split into three classes to account for a possible nonlinear response with bin cut points at 125 *m* and 200 *m*. The most prevalent class (50, 125) was defined as the baseline effect. The resulting model for the delta approach was as follows:

$$\begin{aligned} Y &\sim \mathrm{DG}(\pi, \alpha, \beta) \\ \log\left(\frac{\pi}{1-\pi}\right) &= \zeta + \phi V + \kappa_{\mathrm{Depth}} \\ \frac{\alpha}{\beta} &= \exp(\gamma + \delta V + \eta_{\mathrm{Depth}}) \end{aligned} \qquad \text{eqn 11}$$

where κ_{Depth} and η_{Depth} account for the depth effect. The depth was incorporated via the Poisson intensity parameter in the CPG (consistent with Lecomte *et al.* 2013) although the effect of covariates can be added to either or both of the CPG parameters (Foster & Bravington 2012). The resulting model was as follows:

$$\begin{aligned} Y &\sim CPG(\lambda V, a, b) \\ \log(\lambda) &= \mu + \tau_{\mathrm{Depth}}. \end{aligned} \qquad \text{eqn 12}$$

Where μ is the intercept and τ_{Depth} denotes the depth effect. The same priors and estimation procedure , as the ones used in the simulation study, were considered for the Bayesian inference of the case study. The fitting ability of the two approaches was compared using the deviance information criterion (DIC) (Spiegelhalter *et al.* 2002). The posterior coefficients of variation $\hat{C}_{\mathrm{V}}$, 90% credible intervals CI , and the posterior medians $\hat{\theta}$ were computed for both approaches.

Results

SIMULATION STUDY

The simulation results of the data set generated with parameters set (200,2,0,001) are presented in this section. The three other data sets generated with the sets of parameters (200,2,0,0005), (20,2,0,001) and (200,2;0,15) are provided in Tables S1–S3 as their results are very similar to those of the first data set.

When sampling volume variability was small ($C_{\mathrm{V}} < 0{\cdot}8$), the estimates for the three quantities of interest were good and quite similar for both models, with well calibrated $R_{90\%}$, small RMSE and $\hat{C}_{\mathrm{V}}$ (Table 2). It is worth noting that for a log-normally distributed sampling volume with a unit median and variance σ^2_V, the mean is an increasing function of the variance σ^2_V. Consequently, for the data sets with a small variance, the results did not differ much from those obtained for a constant sampling volume equal to 1. As the C_{V} increases, DG estimates of the probability of absence and positive biomass are overestimated, the recovery ratios decrease, and relative uncertainties surrounding estimated parameters based on the $\hat{C}_{\mathrm{V}}$ increase. In contrast, the CPG approach was able to estimate correctly

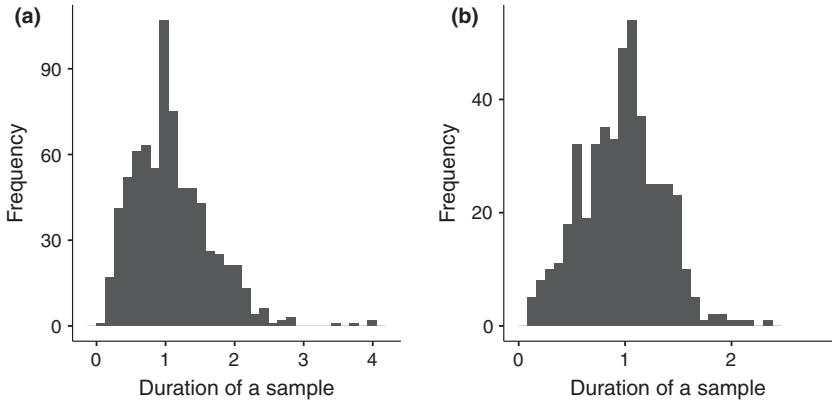

Fig. 1. Histograms of the duration of sampling events of the groundfish commercial catches after rescaling by its mean for the years (a) 2006 and (b) 2009.

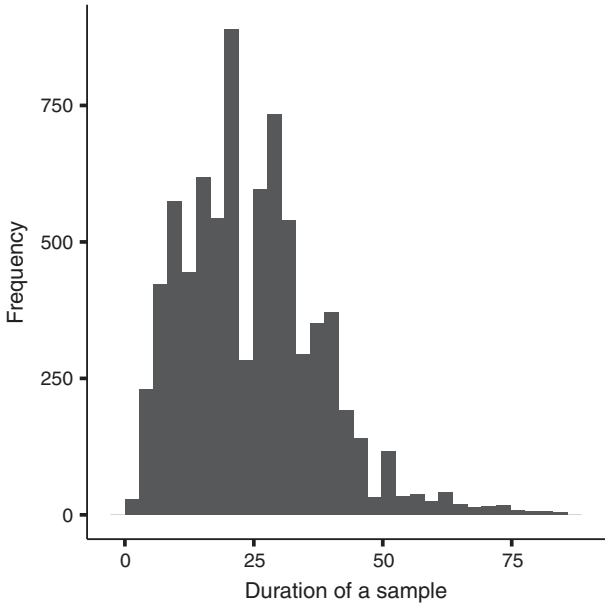

Fig. 2. Histogram of the fishing effort (hours) in the bottom-trawl fisheries of the southern Gulf of St Lawrence (Canada) for Atlantic cod and American plaice, in 1992, the year prior to a moratorium on cod fishing.

the simulated values, with recovery ratios that remained generally correct and constant. Overall, *RMSE* values were lower for the CPG compared to the DG model even when C_V was small. These general patterns remained for different choices of simulated parameters (see Tables S1–S2 in Supporting information).

CASE STUDY: COMMERCIAL FISHERY GROUNDFISH DATA

The probability of absence of dover sole estimated by the two models was high and similar for both years (Table 3). Estimated depth parameters were in accordance between models with depth classes (125, 200) and, (200, 500) having a positive effect on the probability of presence relative to shallow depths, as it was observed with the CPG parameters (Tables 4 and 5). No depth effects were detected with the DG approach for the modelling of the positive biomass (Table 5). In contrast to absence probability, estimates of the overall mean and of the mean positive biomass differed between models for the 2006 data although not the 2009 data (Table 3). Recall that sampling volumes were more variable in 2006 than in 2009.

Results for the ocean perch were similar to those for dover sole. Depth parameters estimates were in accordance between models (Tables 6 and 7). Depth classes (125, 200) and (200, 500) had a positive effect on the presence of the Pacific Ocean perch regarding shallow depths for both years.

Both models similarly estimated a high probability of absence (Table 8), but estimates of the overall mean and of the mean positive biomasses differed dramatically between models for the 2006 data, and to a much less extent for the 2009 data. DIC scores were lower for the CPG than the DG model for both years (Table 9), indicating that the fitting capacity of the CPG model was better than that of the DG model. The CPG model remains a model of choice even in situations where the observed biomass is the sum of a small number of DG-distributed microvolumes biomasses as shown in Table S1, Supporting information.

Discussion

The simulations used in this study allowed for a comparison of two statistical approaches for continuous zero-inflated data by relying on simulated data that mimics the catches of organisms in a uniform habitat with zero-inflation and continuous values of abundance. Based on the simulations, variable sampling volumes were found to produce inference challenges for the DG but not for the CPG distribution. This is consistent with the theoretical arguments we presented concerning the additivity property.

The case study and simulations confirmed that under a variable sampling duration, as it is often encountered in fisheries and other ecological data, the CPG model outperforms the DG overall, providing better fits to data and correct inferences on estimated quantities. The DG model in such situations

Table 2. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{micro} = 200, \beta_{micro} = 2, \pi_{micro} = 0{\cdot}001$)

		$\bar{\hat{\theta}}$		R		RMSE		$\hat{C}_V$	
Volume	θ_{true}	CPG	DG	CPG	DG	CPG	DG	CPG	DG
$C_V = 0{\cdot}1$									
Q	100	**95·83**	95·6	**78**	73	8·05	8·15	0·08	0·08
QP	158·15	**155·42**	154·2	77	**81**	**5·41**	7·62	0·03	0·05
$1-\pi$	0·37	0·38	0·38	77	**88**	**0·03**	0·04	0·08	0·11
$C_V = 0{\cdot}2$									
Q	100	**95·82**	95·56	**85**	84	**7·06**	7·28	0·08	0·08
QP	158·15	**155·41**	155·06	**86**	78	**4·66**	7·69	0·03	0·05
$1-\pi$	0·37	0·38	0·38	85	**87**	0·03	0·03	0·08	0·11
$C_V = 0{\cdot}31$									
Q	100	**96·33**	95·67	87	86	**6·71**	7·19	0·08	0·09
QP	158·15	155·75	**156·04**	88	91	**4·45**	6·49	0·03	0·05
$1-\pi$	0·37	0·38	0·39	87	85	**0·03**	0·04	0·08	0·11
$C_V = 0{\cdot}53$									
Q	100	**96·3**	95·86	88	84	**6·45**	7·58	0·08	0·09
QP	158·15	155·59	**158·8**	87	86	**4·32**	6·43	0·03	0·06
$1-\pi$	0·37	0·38	0·4	87	85	**0·02**	0·04	0·08	0·11
$C_V = 0{\cdot}8$									
Q	100	96·27	**97·28**	84	**86**	**6·37**	8·64	0·07	0·1
QP	158·15	**155·62**	164·61	83	83	**4·23**	10·67	0·03	0·06
$1-\pi$	0·37	**0·38**	0·41	**84**	71	**0·02**	0·05	0·07	0·11
$C_V = 1{\cdot}12$									
Q	100	94·04	**99·77**	76	90	**7·13**	8·34	0·07	0·11
QP	158·15	**154·12**	173·66	**76**	53	**4·74**	17·45	0·03	0·06
$1-\pi$	0·37	**0·39**	0·42	**75**	69	**0·03**	0·06	0·06	0·12
$C_V = 1{\cdot}31$									
Q	100	**95·99**	106·95	**88**	82	**5·88**	12·27	0·07	0·11
QP	158·15	**155·36**	186·94	88	23	**3·97**	31·81	0·03	0·06
$1-\pi$	0·37	**0·38**	0·43	**86**	64	**0·02**	0·06	0·06	0·12

C_V is the coefficient of variation of the simulated sampling volume. θ_{true} is the value used to produce the simulations, $\bar{\hat{\theta}}$ is the posterior median, $R_{90\%}$ is the recovery ratio and should be 90%, RMSE is the root mean squared error and $\hat{C}_V$ is the average estimated coefficient of variation and have to be the lowest. Values in bold denote the best fit.

Table 3. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ for the dover sole sampled in 2006 and 2009

	$\hat{\theta}$		CI		$\hat{C}_V$	
Year	PG	DG	PG	DG	PG	DG
2006						
Q	4·69	7·86	2·95–7·03	4·6-1-2·09	0·27	0·28
QP	100·22	154·09	88·36–113·21	129·38–183·36	0·07	0·1
$1-\pi$	0·95	0·95	0·93–0·97	0·92–0·97	0·01	0·01
2009						
Q	40·9	38·06	31·04–53·54	27·21–52·54	0·17	0·2
QP	154·81	157·76	131·47–184·35	122·24–199·08	0·11	0·14
$1-\pi$	0·74	0·76	0·68–0·79	0·7–0·8	0·05	0·04

$\hat{\theta}$ is the posterior median, CI is the credible interval at 95% and $\hat{C}_V$ is the coefficient of variation.

Table 4. Parameter estimates for the CPG model fitted to the dover sole biomass data sampled in 2006 and 2009

		2006		2009	
Parameter	Term	Mean	SD	Mean	SD
Intercept		−3·047	0·255	−1·187	0·157
Depth	(125,200)	3·13	0·273	−3·401	0·811
Depth	(200,500)	2·985	0·269	0·206	0·226
a		0·528	0·047	0·93	0·138
b		0·005	0·001	0·007	0·001

tends to overestimate mean biomass values, potentially leading to incorrect conclusions, which in the case of fisheries may mean incorrect stock management recommendations. These differences in fitting capacity could be explained by the structure of the CPG model, which can handle variable sampling volumes easily because of the additivity property, whereas the DG approach takes variable sampling volume empirically with the help of a generalized linear model. However, when the sampling volume variability is small, the models performed comparably in the simulation. Fortunately, small sampling volume variability is more the rule than the exception in standardized surveys, and the DG approach therefore remains a valid standard practice in those cases. It is in cases where data do not come from planned surveys, or when data are from two or more surveys with different sampling durations and for which a joint analysis is desired that model choice becomes very

important. This choice can have important ecological and economic consequences. For example, commercial fishery catch-rate data, such as those analysed here for groundfish or the ones exemplified in Fig. 2 for cod, provide the data required to estimate relative abundance indices. These indices form the basis for a large number of stock assessments world-wide, including tuna and cod fisheries that are both highly lucrative and that pose important conservation concerns (e.g. Ahrens (2010); Carruthers *et al*. (2011)). Incorrect inferences drawn from the data are liable to lead to incorrect stock assessment advice and a potential that conservation or economic objectives for a fishery will not be achieved.

In the case study, sampling volume and depth were modelled to affect only the number of patches for the CPG models as, for example, increasing the duration of a sampling event results in an increased number of captured patches. Patch size should vary randomly with respect to changes in sampling volume if a sample is taken in a generally homogeneous habitat. Of course, if increasing sampling volume causes a sample to span more than one area of homogeneous habitat, then both patch number and size can vary in complex ways, and the underlying assumptions of both the CPG and DG could be violated.

Ancelet *et al*. (2010) pointed out a high correlation between the two quantities (number of patches, biomass in one patch) in a special case of the CPG approach. This result suggests that when the CPG distribution is used to model the effect of covariates on the property of interest, such as in generalized linear models (Stefansson 1996; Shono 2008; Zuur *et al*. 2009; Foster & Bravington 2012) or additive models (Zuur *et al*. 2009), it is appropriate to link only one of these two hidden quantities to the explanatory covariates. We suggest that it is most appropriate to model the effect of covariates on the number of patches only, because it tunes both the presence–absence and the quantity of biomass sampled. The parameters are heuristically defined as the number and biomass of patches

Table 5. Parameter estimates for the DG model fitted to the dover sole biomass data sampled in 2006 and 2009

			2006		2009	
Part	Parameter	Term	Mean	SD	Mean	SD
Bernoulli	Intercept		−3·431	0·346	−5·711	0·882
	Volume		0·485	0·176	1·128	0·345
	Depth	(125, 200)	3·459	0·304	3·439	0·799
	Depth	(200, 500)	3·339	0·297	3·753	0·807
Gamma	Intercept		−0·72	0·131	−0·272	0·245
	Volume		0·221	0·091	0·21	0·189
	Depth	(125, 200)	−0·071	0·127	−0·044	0·195
	Depth	(200, 500)	−0·02	0·12	−0·087	0·225

Table 6. Parameter estimates for the CPG model fitted to the pacific ocean perch biomass data sampled in 2006 and 2009

		2006		2009	
Parameter	Term	Mean	SD	Mean	SD
Intercept		−3·434	0·31	−8·452	2·932
Depth	(125, 200)	4·009	0·317	7·194	2·939
Depth	(200, 500)	4·759	0·319	7·849	2·933
a		0·405	0·039	1·568	0·23
b		0·001	0	0·002	0

Table 7. Parameter estimates for the DG model fitted to the pacific ocean perch biomass data sampled in 2006 and 2009

			2006		2009	
Part	Parameter	Term	Mean	SD	Mean	SD
Bernoulli	Intercept		−3·035	0·495	−10·817	3·634
	Volume		−0·077	0·317	1·293	0·344
	Depth	(125, 200)	4·393	0·359	8·356	3·591
	Depth	(200, 500)	7·409	0·647	9·053	3·612
Gamma	Intercept		−0·498	0·101	0·042	0·212
	Volume		0·17	0·075	0·531	0·147
	Depth	(125, 200)	0·014	0·102	−0·306	0·173
	Depth	(200, 500)	0·032	0·098	0·18	0·173

Table 9. Deviance information criterion (DIC) scores related to the DG and CPG models fitted to the data sets of the two species collected in 2006 and 2009 by commercial fisheries

	Perch		Sole	
	2006	2009	2006	2009
DG	7946	2066	4319	1539
CPG	7092	1597	3490	1100

Table 8. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ for the pacific ocean perch sampled in 2006 and 2009. $\hat{\theta}$ is the posterior median, CI is the credible interval at 95% and $\hat{C}_V$ is the coefficient of variation

	$\hat{\theta}$		CI		$\hat{C}_V$	
Year	PG	DG	PG	DG	PG	DG
2006						
Q	17·35	69·32	10·11–27·2	41·06–107·74	0·29	0·3
QP	538·73	1604·82	473·58–610·62	1417·96–1835·28	0·08	0·08
$1-\pi$	0·97	0·96	0·95–0·98	0·94–0·98	0·01	0·01
2009						
Q	281·14	222·35	207·97–358·88	160·41–296·36	0·16	0·18
QP	1136·87	943·78	980·41–1309·96	755·52–1149·25	0·09	0·13
$1-\pi$	0·75	0·76	0·7–0·81	0·71–0·81	0·04	0·04

although these ecological properties are not actually being estimated. Foster & Bravington (2012) used a data set which was composed of biomass and abundance data to explore the relationship between patch size and the size of one typical fish coming from this patch. They showed that the size and the number of patches could have a different relationship to those for the size and number of individual fish. However, such an hypothesis about the size and numbers of patches collected during a sampling event need to be checked. Even if the conjunction of the parameters yields a distribution of biomass values possessing the properties of interest, that is, zero-inflation as well as continuous values with occasional extremes and additivity with respect to variable sampling volume, one must not over interpret an ecological meaning for the individual parameters.

We conclude with practical recommendations arising from this work. When facing zero-inflated data with a constant sampling volume or a sampling volume with a low variability, the DG approach is likely to be understandably preferred by many because of its ease of implementation. However, when working with variable sampling volumes, the analyst should be wary of the DG model. We suggest the CPG structure as a better alternative, even at the cost of some increased complexity of implementation. If not, the simulation study developed in this study shows that, conversely to the CPG, the DG estimates may provide fallacious conclusions, unduly overestimating the biomass quantities.

Acknowledgements

We are indebted to the insightful comments of two anonymous reviewers, which greatly improve the manuscript. We also want to thank an anonymous reviewer for proposing the use of the sampling volumes as a covariate in both parts of the DG approach, which allows for a fair comparison on a more balanced basis.

References

Ahrens, R. (2010) *Global Analysis of Apparent Trends in Abundance and Recruitment of Large Tunas and Billfishes Inferred from Japanese Longline Catch and Effort Data*. PhD thesis. The University of British Columbia, Vancouver, BC.

Ancelet, S., Etienne, M.-P., Benoît, H. P. & Parent, E. (2010) Modelling spatial zeroinated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, **17**, 347–376.

Candy, S. (2004) Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, **11**, 59–80.

Carruthers, T.R., Ahrens, R.N., McAllister, M.K. & Walters, C.J. (2011) Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices. *Fisheries Research*, **109**, 157–167.

Foster, S.D. & Bravington, M.V. (2012) A PoissonGamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, **19**, 1–20.

Jorgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49**, 127–162.

Lecomte, J.B., Benoît, H.P., Etienne, M.P., Bel, L. & Parent, E. (2013) Modeling the habitat associations and spatial distribution of benthic macroinvertebrates, A hierarchical Bayesian model for zeroinated biomass data. *Ecological Modelling*, **265**, 74–84.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A. & Possingham, H. (2005) Zero tolerance ecology, improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.

Maunder, M.N. & Punt, A.E. (2004) Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, **70**, 141–159.

Ntzoufras, I. (2011) *Bayesian Modeling using WinBUGS, volume 698*. Wiley, Hoboken, NJ.

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. *Fisheries Research*, **70**, 275–297.

Pennington, M. (1996) Estimating the mean and variance from highly skewed marine data. *Fishery Bulletin*, **94**, 498–505.

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

Shono, H. (2008) Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, **93**, 154–162.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal stsistical Society: Series B (Statistical Methodology)*.**64**, 583–639.

Stefansson, G. (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**, 577–588.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R. Statistics for Biology and Health*. Springer, New York.

Received 27 August 2013; accepted 27 September 2013
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{micro} = 200, \beta_{micro} = 2, \pi_{micro} = 0{\cdot}0005$).

Table S2. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{micro} = 20, \beta_{micro} = 2, \pi_{micro} = 0{\cdot}001$).

Table S3. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{micro} = 200, \beta_{micro} = 2, \pi_{micro} = 0{\cdot}15$).