Title: K-aggregated transformation of discrete distributions improves modeling count data with excess ones

Can Zhou

(Corresponding author)

Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060

Phone: 1-979-473-9124

Email: eidotog@gmail.com

Yan Jiao

Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060

Joan Browder

Southeast Fisheries Science Center, National Oceanic and Atmospheric Administration, Miami, Florida, 33149

Abstract

The excess one pattern in count data has been documented in ecology but it has not been explicitly modeled or examined. In this study, we introduce a *k*-aggregated transformation of discrete distributions to better model count data with excess ones in a Bayesian generalized linear model framework and demonstrate its use with two groups of case studies (group 1: seabird bycatch in longline fisheries and Legionnaires disease incidence; group 2: survey abundance of Leadbeater's possum and Frigatebird nesting sites). Group 1 examples have a clear excess one data pattern, and these examples are used to demonstrate the concept of the *k*-aggregation technique. On the other hand, group 2 examples lack a clear excess one pattern, and a modeler may not be motivated enough to use the *k*-aggregation technique in these cases. Nonetheless, *k*-aggregated transformation demonstrated better performance for both groups of examples. In all our case studies, the excess zero pattern co-occurred with an excess one pattern, and the excess zeros were modeled thorough either a zero-inflated or hurdle configuration. The better performance of *k*-aggregated distributions is due to their flexibility of adapting to the relatively high frequency of singletons in the data sets. This new technique has broad applicability and utility in improving modeling count data with potential excess ones.

Keywords: count data; rare event; species distribution; Bayesian; generalized linear model.

Introduction

Count processes frequently occur in ecology, for example, species distribution modeling (Cunningham and Lindenmayer, 2005; Lyashevska et al., 2016; Welsh et al., 1996), catch rate analysis (Aidoo et al., 2015; Lo et al., 1992; Ward et al., 2004) and bycatch studies (Brodziak and Walsh, 2013; Martin et al., 2015; Megalofonou, 2005; Minami et al., 2007). Usually, count observations are either log transformed or analyzed using just a handful of count distributions. O'hara and Kotze (2010) showed the dangers of log-transforming count data to fit into continuous data models and advised against transforming count data in general. The reason that many people opt for data transformation could be the scarcity of count distributions. The Poisson distribution may be the most common count distribution, but it requires the equality of the data's mean and variance, a feature which many field data fail to have. Many datasets have a larger variance than the mean, and the negative binomial and beta-binomial distributions, which build upon the binomial distribution, are often used to model such over-dispersed processes (White and Bennetts, 1996). More recently, the Conway-Maxwell-Poisson (CMP) distribution, which can model both over- and under-dispersion with respect to Poisson, has demonstrated superior performance over traditional distributions, and was recommended for analyzing ecological processes (Lynch et al., 2014). However, these distributions may fail to capture some relevant features present in the data. In the following, we present the *excess one* data pattern arising from studies of rare events that all the above-mentioned distributions failing to represent well, and then introduce a *k*-aggregated modeling technique for this type of count data and provide case studies.

A data set has excess ones when it has more singletons than could be explained by the model at hand. It is a relative measure of the richness of the singletons in the data set in two

senses. First, it is relative to the baseline model used to fit the data set, and one can only subjectively measure the number of excess ones after fitting a model; and secondly, it is relative to the other observations in the data set, as they also influence the expected number of singletons under a specific type of distribution. The following two examples, one from seabird bycatch in longline fisheries (Diaz et al., 2009; Zhou et al., 2019) and one from Legionnaires disease incidence (Xu et al., 2014), exhibit a clear excess one pattern that is evident even before model fitting (Figure 1).

Seabird bycatch in longline fisheries exhibit a strong excess one pattern, i.e., a predominance of singleton seabird bycatches, and this feature has been previously reported in the Hawaii longline tuna fishery and western North Atlantic pelagic longline fishery (Gilman et al., 2016; Li et al., 2012). However, this data pattern was never explicitly modeled or examined to see whether it was well represented in the model. In the western North Atlantic pelagic longline fishery (PLL), the number of seabirds caught in one fishing operation (i.e., set), among positive seabird bycatches, ranges from one to nine, with singletons comprising 35.9% of the total (Figure 1A). Seabirds are less frequently caught together. Counts of instances with 2 to 5 seabirds caught approximately halve with increasing number of seabirds caught. A similar yet more pronounced pattern was found in the seabird bycatch data from the Hawaiian longline fishery, where 72% of the total were caught as singletons (Gilman *et al.*, 2016).

The International Commission for the Conservation of Atlantic Tunas recommended a delta log-normal model for modeling seabird bycatch (Hata, 2006; Li and Jiao, 2013; Lo et al., 1992). The excess one pattern in the bycatch data after log-transformation does not conform to a normal distribution, and we suspect potential bias due to this model-data mismatch. The current version of the seabird bycatch model for the western North Atlantic already adopts this *k*-

aggregated modeling technique, and a recent study linking ecological traits of seabirds to bycatch risk potential also makes use of this modeling technique (Zhou et al., 2019). The new total bycatch estimates from 1992 to 2016 were on average 18.81% higher than the original estimates from a log-normal model (Zhou and Jiao, 2017).

In the second case study, we explore the incidence of Legionnaires disease in Singapore, which show a strong excess one pattern (Tang et al., 2017; Xu et al., 2014). Legionnaires disease is a type of acute pneumonia caused by any type of *Legionella* bacteria, which is found in fresh water (Fraser et al., 1977). The infection rate is relatively low, and after exposure, only between 0.1 and 5.0% of the general population develop the disease (Chartier et al., 2007). The weekly Legionnaire disease count data was reported by the Ministry of Health of Singapore in 2005, and it was previously studied by Tang et al. (2017) and Xu et al. (2014). Out of the records among positive Legionnaires cases, 85.2% of the counts are singletons, and records with multiple reports are less frequent (Figure 1B). Tang et al. (2017) found the Poisson distribution failed to capture this excess one feature and propose to model the excess ones in either a maximum likelihood approach based on expectation-maximization algorithm or a Bayesian approach based on MCMC methods. However, they only explored the Poisson distribution as the baseline model, and it is not clear how generally applicable the modeling strategy is, in other words, "*how common is the excess one pattern?*", an issue also faced by the current study.

Admittedly, the examples of seabird bycatch and Legionnaires disease are rather extreme in their degrees of excess one-ness, and such cases are indeed not common. To demonstrate the general utility of our *k*-aggregated technique on more subtle cases, we introduce a second group of examples. These examples include the count data of Leadbeater's possum (*Gymnobelideus leadbeateri*) in southeastern Australia (Lindenmayer et al., 1991) and the number of Frigatebird

(*Fregata minor* and *F. ariel*) nesting sites in the Coral Sea off north-eastern Australia (Cunningham and Lindenmayer, 2005). Both of these datasets have been previously used to demonstrate the utility of the zero-inflated model in ecology (Cunningham and Lindenmayer, 2005). In the Leadbeater's possum dataset, among survey sites with a positive count, singletons represent only 16.1%, and the count of three individuals is the most common case (21.4%, Figure 1C); in the Frigate nesting sites example, singletons represent 31.9% of the positive counts (Figure 1D). In both these cases, the percentage of singletons in the dataset looked so unexceptional that it certainly did not lead the original authors to investigate further.

This study aims to motivate analysts and introduce a family of discrete distributions to better fit count data with excess ones and demonstrate the methodology using two groups of case studies. The first group of case studies serve to demonstrate the excess one data feature, the rationale of the *k*-aggregated technique, and why it works in improving model performance; the second group of case studies serve to demonstrate the broad applicability of the technique and show a more subtle side of the excess one feature in ecological data.

Materials and methods

Case studies

A brief description of the data and their sources of these case studies used can be found in Table 1. Group 1 exhibits a higher percentage of ones than group 2. For each group, we consider one study with and one without covariate(s). Except for the seabird bycatch study, all the data were extracted from published literature. In the following, we give a brief background of only the seabird bycatch study. Detailed accounts of other case studies can be found in the respective source references listed in Table 1.

The U.S. Atlantic pelagic observer program (POP) is a multi-taxa survey program that records the catch of target species, bycatch of seabird and other incidental taxa, environmental information and gear characteristics of the U.S. western North Atlantic longline fleet (Diaz et al., 2009; Li et al., 2016; Zhou et al., 2019). It targets a coverage of 8% of the fleet fishing effort (Diaz et al., 2009). A total of 16,889 longline operations (set/hauls) from the POP were used in this study. Among these, 78 records had a positive seabird bycatch that totaled 145 seabirds observed bycaught between 1992 and 2015.

Probability distributions for the count process

In this study, we consider 1) three base line distributions, including Poisson, negative binomial and Conway-Maxwell-Poisson (CMP) distributions, 2) the zero-truncated versions of those distributions, and 3) a new class of *k*-aggregated distributions for the count process. Poisson and negative binomial distributions are well known and not described here. As a modification of the Poisson distribution, the zero-truncated Poisson has a probability mass function

$$f_{tp}(y=n)=\frac{\lambda^n e^{-\lambda}}{n!}\cdot\frac{1}{1-e^{-\lambda}},$$

with *n* being positive integers, and in a generalized linear model, a log link function is used

$$\log\lambda = c_c + \mathbf{X}_c\theta_c,$$

where c_c is a constant, $\mathbf{X}_c$ is the covariate matrix, and θ_c is a vector of parameters to estimate. A log-linear relationship was assumed between covariates and the parameters of the distribution for the count process. Two case studies have a covariate in the count model: the number of hooks

147 per gear (numerical variable) was included in the seabird bycatch study; the log transformed
148 number of trees with hollows on site was included in the Leadbeater's possum study.

149 The zero-truncated negative binomial has a probability mass function

150 $$f_{tn}(y=n)=\binom{n+r-1}{n}\cdot(1-p)^r\cdot p^n\cdot\frac{1}{1-(1-p)^r},$$

151 where the shape parameter *r* is constrained to be positive, success probability *p* is modeled by
152 assuming *logit*(*p*) has a linear relationship with the covariates, and $\binom{n+r-1}{n}$ is a binomial
153 coefficient, calculated as $\binom{n+r-1}{n}=\frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)}$, where Γ is the gamma function.

154 The CMP distribution is a generalization of the Poisson distribution, and, with one
155 additional shape parameter, it can model both over-dispersion and under-dispersion (Guikema
156 and Goffelt, 2008; Kadane et al., 2006; Shmueli et al., 2005). In contrast, the negative binomial
157 distribution only models over-dispersion with respect to Poisson. In this study, we used the
158 Guikema and Goffelt (2008) formulation of the CMP

159 $$f_{CMP}(y=n)=\frac{1}{S(\mu,\nu)}\left(\frac{\mu^n}{n!}\right)^\nu,$$

160 $$S(\mu,\nu)=\sum_{i=0}^{\infty}\left(\frac{\mu^i}{i!}\right)^\nu,$$

161 where $S(\mu,\nu)$ is a normalizing constant, $\nu\geq 0$ is the shape parameter, $\mu>0$ is the centering
162 parameter of the CMP distribution and it is assumed that log(μ) has a linear relationship with
163 the covariates. This formulation of the CMP distribution has an infinite summation term $S(\mu,\nu)$

which has no close form solution but can be approximated to any arbitrary precision with a large integer value for max *i* (Guikema and Goffelt, 2008 and references therein). In our computation, we explored using different integers for max *i* to balance the accuracy of the model and the computation time, and we found that using 50 as the maximum of *i* was enough for the current study.

K-aggregated transformation

To model excess ones, we aggregated the first *k* probabilities of an original distribution to represent the probability of a singleton outcome, and the probability mass function is

$$f_k(y=n)=\begin{cases} g(0), & n=0 \\ \sum_{i=0}^{k} g(i+1), & n=1 \\ g(n+k), & n\geq 2 \end{cases},$$

with $k:0,1,2,\ldots$, where g is an original distribution. With the transformed distribution, the probability of a singleton outcome is mapped to the sum of probabilities of positive outcomes of less than or equal to $k+1$ from the original distribution, and the probability of outcomes of larger than $k+1$ is mapped to the probability of an outcome of $n+k$ from the original distribution. The transformed distribution is a valid distribution that integrates to one if the original distribution is a proper distribution. Hereafter, we call this new distribution a *k*-aggregated distribution, of which the original distribution is a special case with $k=0$. Parameter *k* acts as a shape parameter to help adapt the modeled distribution to the singleton outcomes in the data. All the models used in this study along with a brief description for each model are listed in Table 2.

Excess Zeros

All the case studies examined here also exhibit zero-inflation (Lambert, 1992). Either a zero-inflated or hurdle configuration was used to model excess zeros in the count data (Zuur et al., 2009).

In a hurdle configuration, the probability of zero observations and the probability of positive observations are modeled separately. A Bernoulli distribution is used for the probability of a zero observation, and truncated-at-zero discrete distributions are used to model the count process.

In a zero-inflated configuration, the probability of a zero observation and the probability of a positive observation are modeled together. A zero observation could come from either the zero mass or the count process, and a latent variable is used in the model to designate where each data point comes from. Case studies B) Legionnaires disease incidence and D) Frigatebird nesting do not include covariate information, and only the probability of a zero catch (p) was modeled.

For the case studies A) seabird bycatch and C) Leadbeater's possum abundance, a logistic regression with a logit-link was used, in both hurdle and zero-inflated configurations, to relate covariates to the probability of a true zero (p) in the Bernoulli distribution

$$logit(p) = c_b + \mathbf{X}_b\theta_b ,$$

where c_b is a constant, $\mathbf{X}_b$ is the covariate matrix for the binary process, and θ_b is the vector of parameters to be estimated. For case study A), covariates used include geographical coordinates (numerical), season (categorical) and target species (categorical). This set of covariates was based on the significance of those variables in predicting seabird bycatch and the availability of

such information in the data (Li and Jiao, 2013; Winter et al., 2011). For case study C), the covariate used is the log transformed number of trees with hollows on site (numerical), which was selected from a pool of covariates including forest age, slope and tree canopy height (Lindenmayer et al., 1991).

Model fitting and comparison

For each case study, we compared models with *k*-aggregated distributions with either a zero-inflated or a hurdle zero structure and the baseline models. Here, we restricted our search for *k* in the range of 0 to 5, and this range turned out to be enough for our case studies; for other cases, the modeler might want to adjust this search range. Thus, for each case study, we fitted 30 GLM with *k*-aggregated distributions excluding the case of $k = 0$, i.e., 3 baseline models $\times$ 5 *k* values $\times 2$ zero configurations, and 6 GLM with baseline distributions, i.e., 3 baseline models $\times 2$ zero configurations. Wide uniform priors were used exclusively in the model.

Model performance was measured based on deviance information criterion (DIC, Spiegelhalter et al., 2002)

$$DIC = \bar{D} + pD \ ,$$

where deviance *D* is twice the negative log-likelihood, $\bar{D}$ is the posterior mean of the deviance, and pD is an estimate of the effective number of parameters in the model. The model with the minimum DIC is the recommended model, and as a rule of thumb, a < 2 difference in DIC relative to the recommended model suggests substantial evidence for the model, differences between 3 and 7 indicate that the model has considerably less support, whereas a larger than 10 difference indicates that the model is very unlikely (Burnham and Anderson, 2003).

A Bayesian method was used to estimate parameters and select models. We used JAGS 4.0 (Plummer, 2003) in the statistical program R 3.2.5 (R Development Core Team, 2016). All the functionalities explored in this paper have been implemented in an easy to use R package *konez* hosted on C.Z.'s GitHub repository. See Appendix for a code example, and for more information, please go to https://hvoltbb.github.io/konez/.

Simulation studies

We further conducted a simulation study based on case study C) Leadbeater's possum abundance to examine the effect of sample size on the selection of *k* and model parameter estimates. We chose this case study for further analysis because this dataset includes covariates and it is based on an animal abundance survey. Most ecologists would find this type of datasets familiar and more relevant to their own work.

Excess one pattern often links to rare events in biology and the corresponding sample sizes are often small as in case studies B) and C) here. In addition to the original sample size (151), we included scenarios with $\times 2$ and $\times 4$ the original sample size, i.e., with 300 and 600 samples. To investigate the effect of sample size on model selection, we generated 500 random datasets based on the selected model, i.e., zero-inflated *k*-aggregated negative binomial with $k = 1$ (Table 3), used a range of models with *k* from 0 to 5 to fit the datasets, and conducted model selection based on DIC. We recorded the number of datasets where the data generating model was selected and the percentage of simulations where the DGM has a $\Delta DIC \leq 2$. To investigate the effect of sample size on model parameter estimates, we generated 500 random samples from the selected model for each sample size scenario, fitted a model with the correct *k* value to each random sample. We checked whether the true parameter lay between the 95% credible interval of the fitted model.

Results

Model selection for 4 case studies

For all our case studies, the selected model based on DIC had a *k*-aggregated distribution (Table 3). The reduction in DIC from the respective best baseline model was between 0.6 and 4.7, indicating a moderate to significant improvement in model performance. The improvement was significant in case studies A) seabird bycatch, B) Legionnaires disease and D) frigatebird nesting sites ($\Delta DIC > 2$). Three case studies A), B, and D) selected a *k*-aggregated CMP model, and only C) Leadbeater's possum abundance selected a *k*-aggregated negative binomial model. The selected value of *k* is relatively small, i.e., $k = 1, 2$, considering our search range ($0 \le k \le 5$).

For each case study, many of the *k*-aggregated models have comparable fit with respect to the selected model (Table 3), i.e., having a less than 2 ΔDIC. Cases A) and C) that included covariate information have, respectively, 8 and 9 models with comparable fit out of the total of 30 candidate models ($k > 1$), and case studies B) and D) that did not include any covariate information have more than twice the number of comparable models. This suggests the importance of covariates in the selection of parameter *k*. On the other hand, none of the baseline models have comparable fit with respect to the selected model as can be seen from case studies A), B) and D). For case study C), two out of six baseline models show comparable fit with respect to the selected model, and both models use the negative binomial distribution but with different zero configurations.

How zero catches were modeled had little effect on model performance for these datasets. Three of the selected models for the 4 cases have a zero-inflated configuration, and one has a hurdle configuration (Table 3). There is no clear relation of the zero configuration between the selected model and the best baseline model. Incidentally, all four best baseline models use the

hurdle configuration. Models with the same type of distribution for the count process but with different assumptions on how zero observations were modeled produced comparable model fit for all our case studies (Supplementary material).

Simulation study of the Leadbeater's possum abundance

With the original sample size, the data generating model (DGM) (*k=1*) was selected 47.20% of the time, and the baseline model was incorrectly selected 26% of the time (Table 4). When a model other than the DGM was selected, the incorrectly selected models have comparable performance with the DGM more than 80% of the time when the original sample size was used. In addition, 95% credible interval estimates have excellent coverage in the simulation study when the original sample size was used (Table 5).

Increasing sample size improved the probability of selecting the correct model for this case study (Table 4). With $\times 4$ the sample size, the DGM was selected 66.80% of the time, and the support for the baseline model almost halved. This trend can also be seen from the shift in concentration of the distribution of the selected *k* over the range of 0 to 5. With increasing sample size, the support for the correct model is increasing. The percentage of runs with $\Delta DIC \leq 2$ only increases slightly when the sample size are about 2 and 4 times the original sample size. The 95% credible interval estimates still have excellent coverage (Table 5).

Discussion

In this study, we introduce a novel method to transform discrete distributions to better fit excess ones in count data within the generalized linear model framework and demonstrated its use in four case studies in ecology using Bayesian methods. In all the case studies, the excess zero pattern co-occurred with excess ones, and the excess zeros were modeled independently through either a zero-inflated or hurdle configuration. The use of *k*-aggregated transformation

lead to moderate to significant improvement in model fitting for our case studies. The better performance of *k*-aggregated distributions is due to the flexibility of adapting to the relatively high frequency of singletons in the dataset. This new technique has broad applicability and utility in improving model fit of count data with potential excess ones.

As suggested by our case studies, the existence of excess-ones in the dataset is a subtle issue, and their presence in ecological data of rare events may be more prevalent than we thought. The seabird bycatch case study is an extreme example of excess ones. This excess-oneness is not unique to the U.S. Atlantic PLL fishery; a large number of singleton bycatches of seabirds are also observed in the Hawaiian PLL fishery (Gilman et al., 2016). However, before conducting a proper analysis on that dataset, we cannot claim if those ones are in excess or not, because excess-oneness is a relative measure. Excess ones may be a common feature of PLL seabird bycatch and even of rare or non-targeted species in general. The other three case studies were also concerning incidence of rare events, whether it is a rare disease or rare species (Cunningham and Lindenmayer, 2005; Tang et al., 2017). The *k*-aggregated distributions also have better performance for a rare spider dataset in Hong Kong presented in the introductory book on zero-inflated models by Zuur and Ieno (2016). This example was not presented in this paper for brevity and interested readers can easily perform the relevant analysis using the provided R package *konez*.

Another feature of this paper is the provision of an easy to use R package *konez* to facilitate the adoption of the *k*-aggregated technique. For example, to perform a model selection of a series of *k*-aggregated models with either a hurdle or zero-inflation configuration requires only one line of code. Please refer to the appendix for an example on how to perform the model selection procedure on the Leadbeater's possum abundance dataset with *konez*. The case studies

318 included in this paper may not represent all potential cases of rare ecological events, but both the
319 simulation study and the 4 case studies demonstrated the potential wide applicability of the *k*-
320 aggregated approach. As more and more ecologists report the analysis of their own datasets with
321 *k*-aggregated distributions, possibly using *konez*, we can better understand the real prevalence of
322 excess-ones in the count data.

323 The *k*-aggregated modeling technique extends existing modeling strategies for rare
324 events. It is a generalization of the baseline count distributions, because with $k = 0$, it reduces to
325 its baseline distribution; it is versatile, because it can be applied to virtually every count
326 distributions; it is an add-on refinement to both the hurdle and zero-inflation models (Lambert,
327 1992; Welsh et al., 1996). Both hurdle and zero-inflated models enjoy extensive usage in
328 modeling literature, including modeling species abundance (Wenger and Freeman, 2008), fishery
329 bycatch (Minami et al., 2007) and catch per unit effort standardization (Shono, 2008; Zhou et al.,
330 2016). All four case studies explored in this paper exhibit excess zeros, and the datasets were
331 originally modeled either through a hurdle or zero-inflated model, in the case of the seabird
332 bycatch example, a delta log-normal model was previously used (Li et al., 2016), which is a
333 special form of hurdle model. The *k*-aggregated modeling technique can be applied along with
334 existing hurdle or zero-inflated structures. As shown in the paper, the *k*-aggregated models out
335 performed all those original models. In particular, the *k*-aggregated Poisson model presented in
336 this paper is similar to the zero-and-one-inflated Poisson models (Tang et al., 2017), and the two
337 models are equivalent when there are no covariates. This paper extends the idea presented in
338 Tang et al. (2017) by generalizing the technique to more count distributions, and including the fit
339 of covariate information through generalized linear models.

The *k*-aggregated technique has broad applicability and utility in improving model fit of count data with potential excess ones. Our work would be useful to fishery and other biological scientists working with count data of rare events. In addition, an R package is provided for researchers to apply a quick analysis of their own dataset using the *k*-aggregated technique.

Acknowledgements

Data used for the seabird bycatch case study were U.S. Atlantic PLL logbook data and POP data from the Fisheries Statistics Division of the Southeast Fisheries Science Center, NOAA National Marine Fisheries Service, Miami, Florida. This research was supported by a contract for Modeling Pelagic Longline Seabird Bycatch awarded to Y. Jiao by the NOAA, NMFS Southeast Fisheries Science Center (SEFSC), under a grant from the NOAA Fisheries' National Seabird Program.

References

Aidoo, E.N., Mueller, U., Goovaerts, P., Hyndes, G.A., 2015. Evaluation of geostatistical estimators and their applicability to characterise the spatial patterns of recreational fishing catch rates. Fisheries research 168, 20-32.

Brodziak, J., Walsh, W.A., 2013. Model selection and multimodel inference for standardizing catch rates of bycatch species: a case study of oceanic whitetip shark in the Hawaii-based longline fishery. Canadian Journal of Fisheries and Aquatic Sciences 70, 1723-1740.

Burnham, K.P., Anderson, D.R., 2003. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Chartier, Y., Lee, J.V., Pond, K., Surman-Lee, S., 2007. Legionella and the prevention of legionellosis. World Health Organization.

Cunningham, R.B., Lindenmayer, D.B., 2005. Modeling count data of rare species: some statistical issues. Ecology 86, 1135-1142.

Diaz, G.A., Beerkircher, L.R., Restrepo, V.R., 2009. Description of the US pelagic observer program (POP). Collect. Vol. Sci. Pap. ICCAT 64, 2415-2426.

Fraser, D.W., Tsai, T.R., Orenstein, W., Parkin, W.E., Beecham, H.J., Sharrar, R.G., Harris, J., Mallison, G.F., Martin, S.M., McDade, J.E., 1977. Legionnaires' disease: description of an epidemic of pneumonia. New England Journal of Medicine 297, 1189-1197.

Gilman, E., Chaloupka, M., Peschon, J., Ellgen, S., 2016. Risk Factors for Seabird Bycatch in a Pelagic Longline Tuna Fishery. PloS one 11, e0155477.

Guikema, S.D., Goffelt, J.P., 2008. A flexible count data regression model for risk analysis. Risk analysis 28, 213-223.

Hata, D., 2006. Incidental captures of seabirds in the US Atlantic pelagic longline fishery, 1986–2005. Unpublished report, National Marine Fisheries Service, Southeast Fisheries Science Center.
Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. Bayesian analysis 1, 363-374.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.
Li, Y., Browder, J.A., Jiao, Y., 2012. Hook effects on seabird bycatch in the United States Atlantic pelagic longline fishery. Bulletin of Marine Science 88, 559-569.
Li, Y., Jiao, Y., 2013. Modeling seabird bycatch in the US Atlantic pelagic longline fishery: Fixed year effect versus random year effect. Ecological modelling 260, 36-41.
Li, Y., Jiao, Y., Browder, J.A., 2016. Assessment of seabird bycatch in the US Atlantic pelagic longline fishery, with an extra exploration on modeling spatial variation. ICES Journal of Marine Science 73, 2687-2694.
Lindenmayer, D., Nix, H., McMahon, J., Hutchinson, M., Tanton, M., 1991. The conservation of Leadbeater's possum, Gymnobelideus leadbeateri (McCoy): a case study of the use of bioclimatic modelling. Journal of Biogeography, 371-383.
Lo, N.C.-h., Jacobson, L.D., Squire, J.L., 1992. Indices of relative abundance from fish spotter data based on Delta-Lognornial Models. Canadian Journal of Fisheries and Aquatic Sciences 49, 2515-2526.
Lyashevska, O., Brus, D.J., Meer, J., 2016. Mapping species abundance by a spatial zero‐inflated Poisson model: a case study in the Wadden Sea, the Netherlands. Ecology and evolution 6, 532-543.
Lynch, H.J., Thorson, J.T., Shelton, A.O., 2014. Dealing with under‐and over‐dispersed count data in life history, spatial, and community ecology. Ecology 95, 3173-3180.
Martin, S.L., Stohs, S.M., Moore, J.E., 2015. Bayesian inference and assessment for rare‐event bycatch in marine fisheries: a drift gillnet fishery case study. Ecological Applications 25, 416-429.
Megalofonou, P., 2005. Incidental catch and estimated discards of pelagic sharks from the swordfish and tuna fisheries in the Mediterranean Sea. Fishery Bulletin 103, 620-634.
Minami, M., Lennert-Cody, C.E., Gao, W., Roman-Verdesoto, M., 2007. Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. Fisheries Research 84, 210-221.
O’hara, R.B., Kotze, D.J., 2010. Do not log‐transform count data. Methods in Ecology and Evolution 1, 118-122.
Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, Proceedings of the 3rd international workshop on distributed statistical computing. Vienna, Austria, p. 125.
R Development Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 127-142.
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., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583-639.
Tang, Y., Liu, W., Xu, A., 2017. Statistical inference for zero-and-one-inflated poisson models. Statistical Theory and Related Fields 1, 216-226.
Ward, P., Myers, R.A., Blanchard, W., 2004. Fish lost at sea: the effect of soak time on pelagic longline catches. Fishery Bulletin 102, 179-195.
Welsh, A.H., Cunningham, R.B., Donnelly, C., Lindenmayer, D.B., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling 88, 297-308.
Wenger, S.J., Freeman, M.C., 2008. Estimating species occurrence, abundance, and detection probability using zero - inflated distributions. Ecology 89, 2953-2959.
White, G.C., Bennetts, R.E., 1996. Analysis of frequency count data using the negative binomial distribution. Ecology 77, 2549-2557.
Winter, A., Jiao, Y., Browder, J.A., 2011. Modeling low rates of seabird bycatch in the US Atlantic longline fishery. Waterbirds 34, 289-303.
Xu, H.-Y., Xie, M., Goh, T.N., 2014. Objective Bayes analysis of zero-inflated Poisson distribution with application to healthcare data. IIE Transactions 46, 843-852.
Zhou, C., Fujiwara, M., Grant, W.E., 2016. Finding regulation among seemingly unregulated populations: a practical framework for analyzing multivariate population time series for their interactions. Environmental and ecological statistics 23, 181-204.
Zhou, C., Jiao, Y., 2017. Estimated seabird bycatch in the U.S. Atlantic pelagic longine fishery during 1992-2016 based on observer and logbook data. The Pelagic Longline Observer Program, Southeast Fisheries Science Center, Miami, FL.
Zhou, C., Jiao, Y., Joan, B., 2019. Seabird bycatch vulnerability to pelagic longline fisheries: ecological traits matter. Aquatic Conservation: Marine and Freshwater Ecosystems (In press).
Zuur, A.F., Ieno, E.N., 2016. Beginner's guide to zero-inflated models with R. Highland Statistics Limited Newburgh.
Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Zero-truncated and zero-inflated models for count data, Mixed effects models and extensions in ecology with R. Springer, pp. 261-293.

451 Table 1 Case studies used in this study.

Group #	Name	Description	Source	Percentage of ones among positive records	Includes Covariate(s)?
1	A) Seabird bycatch	Observed seabird bycatch of U.S. western North Atlantic pelagic longline fleet from 1992 to 2016. A total of 16,889 fishing operations were used, and 78 of them have incidentally caught one or more seabird.	Raw data from the U.S. Atlantic pelagic observer program (Li et al., 2016; Zhou et al., 2019)	35.9	Yes
1	B) Legionnaires disease reports	Weekly Legionnaires disease count in Singapore in 2005 reported to the Ministry of Health of Singapore. The dataset consists of 63 records, 27 of which reported one or more disease report.	Extracted from Xu et al. (2014)	85.2	No
2	C) Leadbeater's possum abundance	Leadbeater's possum is an endangered species only found in the montane ash forests of the Central Highlands of Victoria, Australia. The animal count of 151 survey sites were used in this study.	Extracted from Cunningham and Lindenmayer (2005)	16.1	Yes
2	D) Frigatebird nesting sites	Survey counts of nests of Frigatebird, *F. ariel* and *F. minor*, in Coringa-Herald National Nature Reserve on North East Herald Island, Australia. The number of nests on 236 10x10 m quadrats were used in this study	Extracted from Cunningham and Lindenmayer (2005)	31.9	No

452

453

454 Table 2 Summary of all the models used in this study. With $k = 1$, the corresponding model is

455 the same as the commonly used hurdle model or zero-inflated model.

Baseline Model	Category	Description
Poisson	Hurdle	Hurdle *k*-aggregated zero-truncated Poisson GLM with *k*=0, 1, 2, …, 5
Negative binomial	-	Hurdle *k*-aggregated zero-truncated negative binomial GLM with *k*=0, 1, 2, …, 5
CMP	-	Hurdle *k*-aggregated zero-truncated CMP GLM with *k*=0, 1, 2, …, 5
Poisson	Zero-inflation	Zero-inflated *k*-aggregated Poisson GLM with *k*=0, 1, 2, …, 5
Negative binomial	-	Zero-inflated *k*-aggregated negative binomial GLM with *k*=0, 1, 2, …, 5
CMP	-	Zero-inflated *k*-aggregated CMP GLM with *k*=0, 1, 2, …, 5

456

457 Table 3 Model selection results of four case studies. For the abbreviations used in this table,
458 CMP stands for Conway-Maxwell-Poisson distribution, NB stands for negative binomial
459 distribution, H stands for hurdle configuration of excess zeros, and ZI stands for the zero-inflated
460 configuration of excess zeros.

Case study	Selected model (base line dist., k value, zero config.)	Best baseline model, zero config. (ΔDIC)	Number of models with ΔDIC $\leq$ 2 among candidate models	
			K-aggregated model	Baseline models
A) Seabird bycatch	CMP, k=2, H	NB, H (4.7)	8/30	0/6
B) Legionnaires disease reports	CMP, k=1, ZI	CMP, H (2.5)	17/30	0/6
C) Leadbeater's possum abundance	NB, k=1, ZI	NB, H (0.6)	9/30	2/6
D) Frigatebird nesting sites	CMP, k=1, ZI	CMP, H (3.3)	20/30	0/6

461

462 Table 4 Model selection result of the 500 simulated replicates for each of three different sample
463 sizes (original, $\times 2$ and $\times 4$) based on the selected model (*k*-aggregated negative binomial with
464 $k = 1$ and zero-inflation) of the Leadbeater's possum abundance dataset. DGM stands for data
465 generating model.

Sample size	Number of selected models with *k*						Percentage of replicates when the DGM is selected	Percentage of simulations when the DGM has a Δ DIC $\leq$ 2
	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$		
$\times 1$	130	236	99	26	5	4	47.20%	80.40%
$\times 2$	108	262	103	24	3	0	52.40%	85.00%
$\times 4$	71	334	86	9	0	0	66.80%	86.20%

466

Table 5 Coverage percentage of 95% credible intervals for each of three different sample sizes (original, $\times 2$ and $\times 4$) based on the selected model (*k*-aggregated negative binomial with $k = 1$ and zero-inflation) of the Leadbeater's possum abundance dataset

Sample size	Percentage of 95% credible intervals that covers the true value				
	b_1	c_1	b_2	c_2	r
$\times 1$	95.0%	93.4%	94.2%	94.0%	96.8%
$\times 2$	96.0%	95.4%	93.8%	94.6%	96.0%
$\times 4$	95.2%	95.2%	95.6%	94.4%	96.8%

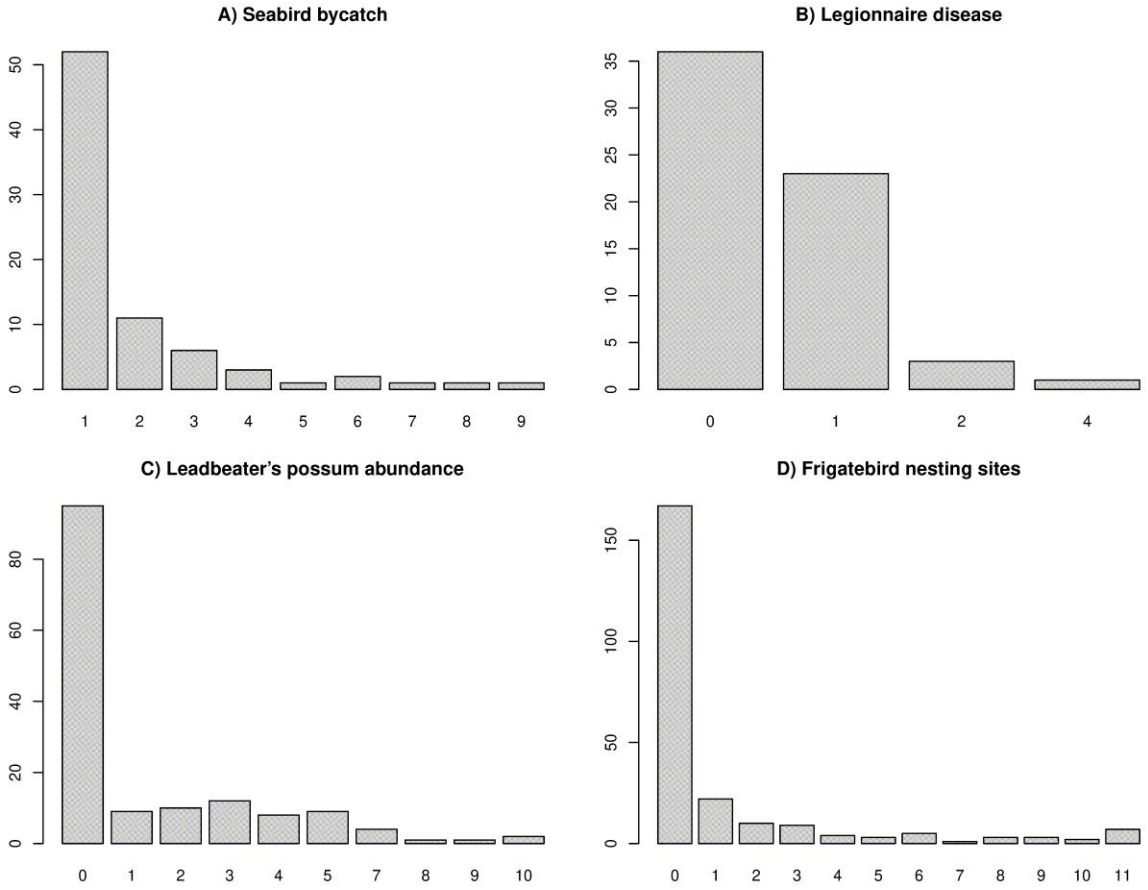

Figure 1 Count data from two groups of case studies. Group #1 examples include A) seabird bycatch in pelagic longline fishery and B) Legionnaire disease reports [see Xu et al. (2014) for more details]; Group #2 examples include C) survey abundance of Leadbeater's possum [see Lindenmayer et al. (1991) for further details] and D) Frigatebird nesting sites [see Cunningham and Lindenmayer (2005) for further details]. Except the seabird bycatch example, all relevant data including covariate(s) were extracted from the cited literature. In panel A, the frequency of zero counts of seabird bycatch were not plotted, which consists of more than 99% of all the counts; in panel B, the count of three has zero occurrence in the dataset, and it was not plotted.