A novel method for quantifying overdispersion in count data and its application to farmland birds

Article *in* Ibis · December 2016

DOI: 10.1111/ibi.12450

Barry J. McMahon
University College Dublin
112 PUBLICATIONS **2,007** CITATIONS

Helen Sheridan
University College Dublin
63 PUBLICATIONS **795** CITATIONS

Andrew C Parnell
Maynooth University
164 PUBLICATIONS **13,541** CITATIONS

IBIS
BOU international journal of avian science

Ibis (2017), **159**, 406–414 doi: 10.1111/ibi.12450

A novel method for quantifying overdispersion in count data and its application to farmland birds

BARRY J. MCMAHON,[1,*] GORDON PURVIS,[1] HELEN SHERIDAN,[1] GAVIN M. SIRIWARDENA[2] & ANDREW C. PARNELL[3]

[1] *UCD School of Agriculture & Food Science, University College Dublin, Belfield, Dublin 4, Ireland*
[2] *British Trust for Ornithology, The Nunnery, Thetford, Norfolk IP24 2PU, UK*
[3] *UCD School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin, Belfield Campus, Dublin 4, Ireland*

The statistical modelling of count data permeates the discipline of ecology. Such data often exhibit overdispersion compared with a standard Poisson distribution, so that the variance of the counts is greater than that of the mean. Whereas modelling to reveal the effects of explanatory variables on the mean is commonplace, overdispersion is generally regarded as a nuisance parameter to be accounted for and subsequently ignored. Instead, we propose a method that models the overdispersion as a biologically interesting property of a data set and show how novel inference is provided as a result. We adapted the double hierarchical generalized linear model approach to create an easily extendible model structure that quantifies the influence of explanatory variables on the overdispersion of count data, and apply it to farmland birds. These data were from a study within Irish agricultural ecosystems, in which total bird species abundance and the abundance of farmland indicator species were compared on dairy and non-dairy farms in the winter and breeding seasons. In general, overdispersion in bird counts was greater on dairy farms than on non-dairy farms, and for total bird numbers, overdispersion was greatest on dairy farms in winter. Our code is fitted using the Bayesian package *Rstan*, and we make all code and data available in a GitHub repository. Within a Bayesian framework, this approach facilitates a meaningful quantification of the effects of categorical explanatory variables on any response variable with a tendency to overdispersion that has a meaningful biological or ecological explanation.

Keywords: abundance, agricultural systems, Bayesian framework, ecological data.

Count data exhibiting overdispersion are a common occurrence in ecological statistical modelling (e.g. O'Hara & Kotze 2010, Harrison 2015). Such data arise when the variance of the counts is larger than the mean, as opposed to the standard Poisson probability distribution, which requires their equality. Many reasons have been proposed for the existence of overdispersion (Lindén & Mäntyniemi 2011, McMahon *et al.* 2013a). When explanatory variables are available, it is common simply to quantify their effect on the mean. However, the overdispersion may also reflect biologically interesting patterns, such as that of flocking (Lindén & Mäntyniemi 2011), so its drivers warrant investigation. Flocking may be a mechanism to enable bird species to source food more efficiently or to increase protection against predation. We provide an easily extendible route towards achieving greater insight into the ecological meaning of excess variance. We adapt the double hierarchical generalized linear model approach of Lee and Nelder (2006) which, to our knowledge, has yet to be applied in empirical ecological studies.

Although Poisson is the most commonly used probability distribution for counts, many extensions and alternatives have been proposed to model such data in the presence of overdispersion,

[*] Corresponding author.
Email: barry.mcmahon@ucd.ie

Associate Editor: Lluis Brotons.

such as quasi-Poisson models (Wedderburn 1974), random-effects models (Bolker *et al.* 2008) and negative-binomial models (McCullagh & Nelder 1989). Indeed, the negative binomial model that we use can be seen as a marginalized gamma-distributed random effects model applied to Poisson data (Agresti 2013). Further intense debate has focused on whether a count distribution is valid for overdispersed data at all, and whether it might be better replaced by appropriate data transformation (e.g. via the square root or log) and use of a more general model (O'Hara & Kotze 2010, Ives 2015). Our proposed approach neatly sidesteps this question, as it accounts for, and allows greater control of, the mean/variance relationship.

We adopt a Bayesian approach to statistical inference. This has the advantages of being more easily extendible, creating more coherent and understandable probability distributions for the user and allowing for prior information to be incorporated in the analysis. However, our model is agnostic to the inferential paradigm used, e.g. maximum likelihood estimates. We fitted our models using the Hamiltonian Monte Carlo package (RStan Stan Development Team 2015), as this provides fast access to the full posterior distribution of the parameters. We provide all our code in a GitHub repository (https://github.com/andrewcparnell/birds_od).

Our particular focus was on the modelling of overdispersion, i.e. the incorporation of explanatory variables that allow the determination of the underlying causes of excess variance. This has not been the focus of much previous research (although see Lindén & Mäntyniemi 2011), as many studies treat any parameters associated with overdispersion as constant across observations. In contrast, our approach, with a re-parameterization of the negative binomial distribution, allowed us to examine overdispersion from a range of overdispersion relationships for which the quasi-Poisson and the negative binomial are special cases.

Rather than treat overdispersion simply as a nuisance parameter, and correct for its bias in statistical tests, it has been suggested that an appropriate approach would be to consider the biological mechanisms that might plausibly underpin the phenomenon, such as flocking in birds, and this logic might usefully be incorporated into the modelling process. If so, the data should follow a negative binomial distribution, with the mean–variance relationship governed by an additional parameter that quantifies the flocking or other aggregation behaviour on the response variable (McMahon *et al.* 2013a).

We tested the potential of the method with a data set on farmland birds in Ireland, where variation in both mean abundance and flocking might be expected between farm systems and seasons. Our approach allows for the calculation of estimates of overdispersion and quantitatively evaluates the influence of explanatory variables, in this case farm system (non-dairy and dairy) and survey season (winter and breeding). We discuss the wider use of this modelling approach in drawing inferences from biological count data.

METHODS

Statistical models for overdispersion

The most common framework to deal with overdispersion in count data is that of generalized linear models (GLMs; McCullagh & Nelder 1989). This approach broadly provides two distributions on which to model overdispersion count responses, namely, the quasi-Poisson and the negative binomial. Perhaps the most important difference between these two distributions for the problem at hand is the behaviour of the variance with respect to the mean. In the quasi-Poisson, the variance is a linear function of the mean, whereas in the negative binomial, the variance is a quadratic function of the mean. In either case, it could be argued that such a strict relationship is overly prescriptive, and further extensions are required explicitly to model the overdispersion.

A body of research has been devoted to comparing and evaluating these two approaches (see Ver Hoef & Boveng 2007 for a review). The GLM paradigm allows for explanatory variables to be included in the model via a link function applied to the mean. However, there is surprisingly little research on the parametric modelling of overdispersion. Lindén and Mäntyniemi (2011) propose a specific new parameterization which bridges the quasi-Poisson/negative binomial divide, but does not allow for explanatory variables to be included that more flexibly account for overdispersion.

A common motivation for the negative binomial distribution is that it arises as a gamma random effect applied to the mean or equivalently variance, the two being mathematically linked in the Poisson distribution. It is thus feasible to create overdispersed counts through the application of random effects to the mean of a Poisson distribution through the link function. However, this approach makes identification of the causal effects hard, especially if there are multiple explanatory variables. Our approach, in contrast, allows for easy examination of the factors that cause variance in excess of the mean.

Concurrently, in GLM modelling of count data with overdispersion, there has been much debate as to whether it is valid to transform the data (e.g. via log or square-root transformations) prior to use of standard linear models (see O'Hara & Kotze 2010, and Ives 2015 for both sides of this argument). One justification for using the linear model approach is the increased control over the mean–variance relationship, which can be entirely independent, whereas, as noted above, using the GLM approach assumes a prescribed relationship between mean and variance. Our new approach relaxes this constraint while remaining within the GLM framework, and so perhaps allows for compromise in this debate.

Modelling approach

We expressed our univariate response variable as counts y_i for observation $i = 1, \ldots n$. In our application study below, there were two potential response variables, although we considered only univariate response models for illustration. Multivariate extensions are addressed in the Discussion. Given the existence of overdispersion in the data, we used a negative binomial distribution for y. We used the alternative parameterization of the negative binomial:

$$y_i \sim \mathrm{NB}(\phi_i,\ \mu_i)$$

where φ_i is the overdispersion parameter and μ_i the mean rate parameter, both constrained to be > 0. The mean and variance of this distribution were:

$$\mathbb{E}(y_i) = \mu_i; \mathrm{Var}(y_i) = \mu_i + \frac{\mu_i^2}{\phi_i}$$

The log-link function was defined so that $\log(\mu_i) = \beta_0 + x_i^T\beta$ was dependent on any explanatory variables x_i through parameters β and to fix $\varphi_i = \varphi$ and a single overdispersion parameter was estimated.

If we wish to model the dependence of the overdispersion parameter on the explanatory variables, it is not helpful to model φ_i through a link function, as the overdispersion is linked to the mean. For example, setting $\log(\phi_i) = \gamma_0 + x_i^T\gamma$ for a new set of parameters γ will yield a complex excess variance relationship that depends on the explanatory variables through both β and γ. This is exactly the case in the GAMLSS package in R (R Development Core Team, 2005), which is widely utilized in ecology (Katsanevakis *et al.* 2010, Kiffner *et al.* 2011, Ourens *et al.* 2014). The interpretation of how the explanatory variables cause the excess variance is extremely difficult using this parameterization.

As an alternative, we modelled the dispersion via the re-parameterization $\theta_i = \mu_i^2/\phi_i$ which gave $\mathrm{Var}(y_i) = \mu_i + \theta_i$. Thus θ_i directly modelled the overdispersion. As it is also necessarily positive, we applied the log-link function to its value and connected it to the explanatory variables, yielding $\log(\theta_i) = \gamma_0 + x_i^T\gamma$. We used this approach to measure separately the effect of the explanatory variables on the mean (via β) and the overdispersion (via γ). As will be seen in the Results, the interpretation of these parameters was clearer than when the traditional parameterization was used.

Our novel parameterization had further subtle benefits. If we introduce the mean as an offset, i.e. $\log(\theta_i) = \gamma_0 + \log(\mu_i)$, a quasi-Poisson type variance–mean relationship is achieved. Alternatively, if we add in a quadratic function of the mean, a negative binomial formulation is achieved. In practice, we found it preferable to do neither of these, and allowed the explanatory variables to guide the behaviour of the variance relationship. Furthermore, it is conceivable that we might choose an alternative to the log-link on θ_i, and so allow for underdispersion. However, for brevity, we do not explore that approach further here.

Finally, we did not perform any model selection in this paper. As our approach is a re-formulation of the negative binomial GLM, it is feasible to compute any of the standard metrics by which models can be compared, including DIC (Spiegelhalter *et al.* 2002) or the more recently introduced WAIC/WBIC (Watanabe 2013). Our goal here is to show how overdispersion can be modelled better through this transformation, rather than explicitly exploring the causal or predictive structure of our data.

The modelling structure can be fully stated hierarchically including potential offsets (o_i) and random effects b_i for the mean and a_i for the overdispersion. We used:

$$y_i \sim \mathrm{NB}(\mu_i^2/\theta_i,\ \mu_i)$$

$$\log(\mu_i) = o_i + \beta_0 + x_i^T\beta + z_i^T b_i$$

$$\log(\theta_i) = o_i + \gamma_0 + x_i^T\gamma + z_i^T a_i$$

These random effects may allow for nesting, and are given normally distributed priors, e.g. $b_i \sim N(0, \sigma_b^2)$ in the univariate case. Substituting in $\theta_i = \mu_i^2/\phi_i$, as detailed above, yields the standard negative binomial formation.

Model fitting

We fitted the models using the Bayesian Monte Carlo package Stan (Stan Development Team 2015). The package works by simulating from the posterior probability distribution of the parameters given the data, using an efficient Hamiltonian implementation of Markov chain Monte Carlo known as the No U-Turn Sampler (NUTS; Hoffman & Gelman 2014). Bayesian modelling is now routine in ecology (McCarthy 2007) and we do not discuss the computational issues further. All the modelling details, including Stan code, data, and R code to check convergence of the fitted model and produce the plots in this paper, are available at the GitHub repository: https://github.com/andrewcparnell/birds_od.

Application study: farmland birds

Data from a study of bird populations within Irish farmland were used to demonstrate our novel approach (McMahon *et al.* 2013b). The rationale behind this was to compare abundance patterns between farm systems (non-dairy and dairy) and survey seasons (winter and breeding), as variations had been observed previously (McMahon *et al.* 2013b).

Site selection

Farm selection for winter and breeding season bird surveys provided a random selection of dairy and non-dairy farms reflecting the proportional incidence of farm types along a north–south gradient of increasing intensity in grassland farming practice in three separate geographical regions of the Republic of Ireland, counties Sligo/Leitrim, Offaly/Laois and Cork (Emerson & Gillmor 1999, Lafferty *et al.* 1999). In total, 40 farms were selected within each region (20 per year in each sampling region), giving a total sample of 120 surveyed farms. For further survey details, see McMahon *et al.* (2012, 2013b).

Bird data

Each farm was surveyed once in the winter season (December–February) and once in the breeding season (April–June). The same surveyor carried out all surveys using a standardized protocol (McMahon *et al.* 2013b). During the breeding season, surveys were carried out between 07.00 and 12.00 h, whereas during the winter season, surveys were carried out between 10.00 and 15.00 h. The mean duration (± sd) of surveys in the winter season was 61 ± 13 min, and in the breeding season 67 ± 18 min. The number, abundance and location of bird species in field boundaries were recorded directly onto site maps, including raptors seen hunting over fields and field boundaries. Other species seen flying overhead, but not interacting with fields or field boundaries, were not recorded. The bird data were collated into total bird species abundance and the abundance of specialist, farmland indicator species (Gregory *et al.* 2004).

Exploratory data analyses

In the winter period, 60 farms were surveyed in 2007–2008, and 59 farms in 2008–2009. During the breeding season, only 42 of the selected farms could be surveyed in 2007 due to generally poor weather conditions, and 59 farms were surveyed in 2008. A breakdown of all farms surveyed by region and farm system is presented in Table 1. All species recorded are presented in Supporting Information Table S1.

A boxplot comparing overall bird abundance and farmland indicator species abundance is shown in

Table 1. The region and system breakdown of the farms utilized in the study.

Region	System	*n*
Cork	Dairy	19
	Non-dairy	20
Offaly/Laois	Dairy	13
	Non-dairy	27
Sligo/Leitrim	Dairy	4
	Non-dairy	36

Figure 1, comparing non-dairy and dairy farms. Higher counts were observed on dairy than on non-dairy farms. The data were strongly overdispersed with, in all cases, the variances being at least an order of magnitude larger than means. Figure 2 shows similar boxplots for counts made in winter and breeding season surveys. Counts were higher in the winter than in the breeding season, although this was less pronounced for farmland indicator species compared with total bird abundance.

As part of an explanatory data analysis, we also looked at other potential explanatory variables, including calendar day of surveys, and a potential offset – the duration of surveys. None of these variables was found to be informative and we did not explore these variables further.

Modelling

Our modelling strategy was to explore how the two response variables, overall bird abundance and farmland indicator species abundance, and the overdispersion in these variables, were affected by a number of explanatory variables, some of which were included in the aforementioned plots. Our explanatory variables included:

- System (non-dairy/dairy): Of key importance was to determine how the counts were affected by non-dairy or dairy farms. Figure 1 suggests that counts were higher on dairy farms.
- Season (winter/breeding): Figure 2 suggests that counts were higher in winter, although this may have arisen through interactions with other variables.
- Region: Three regions were sampled as part of the study (Sligo/Leitrim, Offaly/Laois and Cork). We included this as a random effect and quantified the variability between regions.
- Square: Within each region, a set of 10, 10-km squares were chosen for sampling. This was a nested random effect within region (i.e. there were 10, 10-km squares within each region), therefore square was included to quantify its variability.

The two response variables were overdispersed compared with a standard Poisson distribution and our primary interest was in the causes of the overdispersion. We fitted the models to each response variable in turn with parameters as defined in the Methods section above. The structure of the relationship with the explanatory variables was:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{dairy}_i + \beta_2 \text{winter}_i + \beta_3 \text{dairy}_i \times \text{winter}_i + b_{\text{region}_i} + b_{\text{square}\backslash\text{region}_i}$$

$$\log(\theta_{\text{i}}) = \gamma_0 + \gamma_1 \text{dairy}_i + \gamma_2 \text{winter}_i + \gamma_3 \text{dairy}_i \times \text{winter}_i$$

For the region random effect we defined $b_{\text{region}_\text{i}} \sim N(0, \sigma^2_{\text{region}})$ and for the nested square effect within region we defined $b_{\text{square}\backslash\text{region}_i} \sim N(0, \sigma^2_{\text{square}})$. Our aim was to estimate all of the unknown parameters given the data. The posterior distributions of the γ inform us as to the role of the explanatory variables in the overdispersion of the counts.

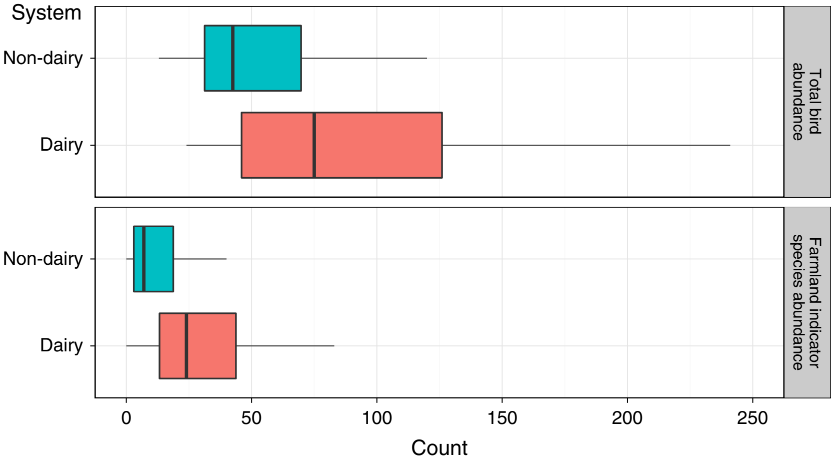

Figure 1. Boxplots representing the model output of the mean and interquartile range along with lines that represent the 95% credible intervals of overall bird abundance and farmland indicator species abundance for non-dairy and dairy farms. [Colour figure can be viewed at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1474-919X]

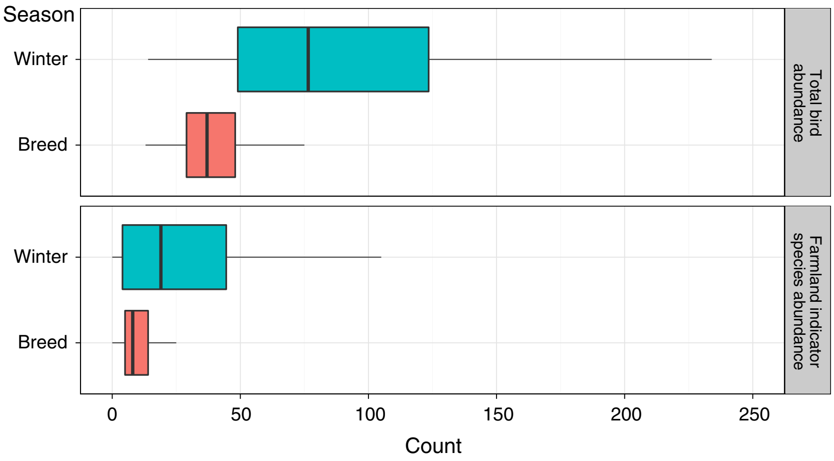

Figure 2. Boxplots representing the model output of the mean and interquartile range along with lines that represent the 95% credible intervals of overall bird abundance and farmland indicator species abundance for winter and breeding season. [Colour figure can be viewed at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1474-919X]

Model comparison, as stated above, was not the main focus of this paper. Rather, our approach allows for simpler interpretation of any given model for overdispersion, despite the fact that the fit might be equivalent to that of a standard (e.g. GAMLSS) model. Further analysis (not shown) using WAIC (Watanabe 2013) indicated that the dairy by winter term in the mean and overdispersion were important components in the model. It is relatively straightforward to extend our code using the loo package (Vehtari *et al.* 2016).

The Bayesian model we fitted requires prior distributions for all parameters. Best practice dictates that informative prior distributions should be used wherever possible, and weakly informative priors to constrain the model fit otherwise. We placed $N(0, 10^2)$ prior distributions on the β and γ parameters. This indicates that most of the explanatory variables are unlikely to influence the counts beyond a value of ± 20 on the log scale, a value that seems uninformative in the absence of further information. For the random effect variance parameters, we used half-Cauchy $t_1(0, 10)$ distributions, which are only weakly informative (Gelman 2006).

RESULTS

Model output

Figure 3 illustrates the posterior distributions of $\exp(\beta_1)$ (mean effect of dairy), $\exp(\beta_2)$ (mean effect of winter) and $\exp(\beta_3)$ (mean effect of the interaction between dairy and winter). The winter season and dairy farm effects increased total bird abundance by a small but positive multiplicative factor (95% credible intervals (CIs) were 0.95–1.59 and 1.02–1.48, respectively). However, the interaction between dairy and winter effects had the greatest effect on observed mean total bird counts (95% CI 1.82–3.71). Note that these are multiplicative effects and should be interpreted accordingly.

For farmland indicator species abundance, both dairy and winter made a clear positive contribution to recorded counts. The 95% CI was 1.93–2.67 and 1.43–2.71, respectively. However, the interaction term was less well quantified and appeared to make a relatively smaller contribution, with 95% CI 0.52–1.48. These individual effects were discernible in exploratory plots (Figs 1 and 2), but the models are of considerably greater help in quantifying the relative scale of these effects, and most especially in the interpretation of the interactions (Fig. 3).

Figure 4 illustrates the posterior distributions of $\exp(\gamma_1)$, the overdispersion effect on dairy, $\exp(\gamma_2)$, the overdispersion effect in winter surveys, and $\exp(\gamma_3)$, the overdispersion effect in interaction between dairy and winter variables. These values are the increase in variance due to the effect in question. The interpretation is thus far simpler due to the change in parameterization. For overall bird abundance, the increase in overdispersion due to the interaction between dairy and winter seemed most pronounced (95% CI 12.17–73.27). For

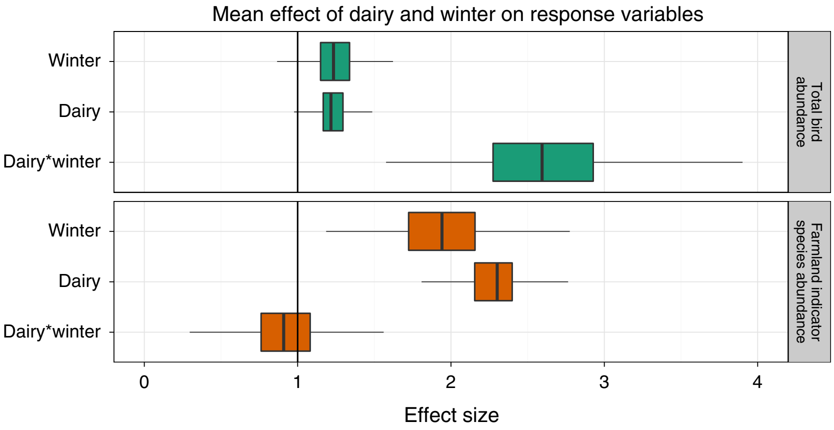

Figure 3. Boxplots representing the model output for the mean effect interquartile range along with lines that represent the 95% credible intervals of farm system, survey season and their interaction on the mean overall bird abundance and farmland indicator species abundance. All but the interaction term for farmland indicator species abundance seem identifiable and well quantified. [Colour figure can be viewed at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1474-919X]

farmland indicator species abundance, the biggest contributor was the effect of dairy farms (95% CI 14.69–55.60). However, these large effects were sometimes poorly quantified.

DISCUSSION

To our knowledge, this is the first time that a statistical modelling approach has been applied that has allowed direct quantification of the influence of explanatory variables on overdispersion in count data. To achieve this, we re-parameterized a GLM using the negative binomial distribution, which facilitates complete separation of effects on the mean and data overdispersion within a Bayesian framework. Several potential extensions of this approach may be considered. One possibility is in multivariate modelling of count data, for which a copula approach might be most feasible (e.g. Nikoloulopoulos & Karlis 2009). Other possibilities stemming from the standard GLM literature include application in the analysis of time series components, spatial models or shrinkage priors for variable selection. A more thorough analysis of our current study data might involve use of the approach in a further exploration of modelling techniques to refine and optimize quantification of explanatory variable effects, including species-specific models and how habitat at various spatial scales might influence abundance and overdispersion. Habitat features at farm and landscape scales, including habitat composition and configuration, could be tested to examine their influence on the abundance and overdispersion of farmland bird communities. In addition, the influence of specific habitat features which are known to influence specific bird species in winter, e.g. cereal stubble and Yellowhammer *Emberiza citrinella* (Gillings *et al.* 2005), could be modelled to examine their influence on overdispersion along with abundance.

In the current application study, overdispersion was most evident in counts of both total bird abundance and farmland indicator species abundance on dairy compared with non-dairy farms. This may be a symptom of greater population aggregation on farms with more intensive stocking rates and nutrient inputs, which are greater on dairy farms (McMahon *et al.* 2010, 2013b). In general, overdispersion in the current data can be interpreted most plausibly as evidence of population aggregation in response to food availability. For total bird abundance, this response is clearly greater in the winter season, when flocking, including the appearance of winter migrant species, occurs. In contrast, the majority of bird species are likely to become more strongly territorial (and therefore more evenly dispersed) in the breeding season. In comparison, the current analysis suggests that the distribution of farmland indicator bird species is less clear, and is possibly impacted by the greater density of food resources on dairy than on non-dairy farms. However, real ecological inference would best be sought from species-specific analyses, rather than those using variables constructed from data aggregated across species, as in this study. Therefore the results here are best

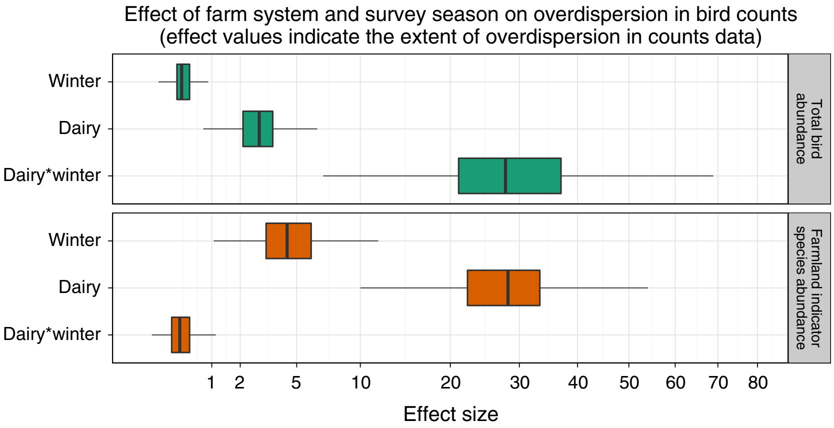

Figure 4. Boxplots representing the model output for the interquartile range along with lines that represent the 95% credible intervals for the overdispersion effects of season and system on overall abundance and farmland indicator species abundance. These are the additional variance of the counts due to individual factors. The *x*-axis is plotted on a square-root scale so as to allow for comparison on the scale of the data (i.e. as counts). For overall abundance, the winter season on non-dairy farms seems to have the smallest effect on overdispersion, whereas in dairy farms and on dairy farms in winter, the overdispersion is increased. For farmland indicator species abundance, both the winter season and dairy farms appear to increase overdispersion. Their interaction effect is perhaps negative but poorly quantified. [Colour figure can be viewed at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1474-919X]

interpreted as proof, example application and adaptation of a previously proposed concept, that of Lee and Nelder (2006).

In the current application of this novel approach to understand better the influence of data dispersion in GLM modelling, added insight is gained to support biologically meaningful interpretation. Application of the approach should be useful in many contexts beyond the analysis of farmland bird data and, within a Bayesian framework, to facilitate more meaningful quantification of the effects of categorical explanatory variables on any response variable with a tendency to overdispersion that has a meaningful biological/ecological explanation. Examples include species that tend to aggregate within habitats, or indeed parasites that congregate on specific hosts (Elston *et al.* 2001, Johnson & Fritz 2014). The quantification and comparison of overdispersion can enable ecologists to understand more about the causes and correlates of aggregation, and usefully inform appropriate environmental management.

We demonstrate an innovative approach to modelling overdispersed ecological data within a flexible Bayesian framework. As with other approaches, e.g. Lindén and Mäntyniemi (2011), understanding the processes causing overdispersion in count data is vital to fit a meaningful model that best describes the biological system involved. Rather than treating overdispersion as a nuisance variable, this approach to the analysis of overdispersed data can potentially add to ecological understanding by quantifying its underlying cause.

Funding was provided by the Irish Department of Agriculture, Fisheries and Food Research Stimulus Fund 2006 (06 0376). The co-operation of farmers who provided sites was greatly appreciated. We thank Tim Carnus for useful discussions during this study and Alison Johnston, Dan Chamberlain, Lluís Brotons and an anonymous reviewer for comments on a previous version of the manuscript.

REFERENCES

Agresti, A. 2013. *Categorical Data Analysis*. Wiley. Ser. Prob. Stat. New York, NY: Wiley-Blackwell.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* **24**: 127–135.

Elston, D., Moss, R., Boulinier, C., Arrowsmith, C. & Lambin, X. 2001. Analysis of aggregation, a worked example: numbers of ticks on Red Grouse chicks. *Parasitology* **122**: 563–569.

Emerson, H.J. & Gillmor, D.A. 1999. The rural environment protection scheme of the Republic of Ireland. *Land Use Policy* **16**: 235–245.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**: 515–534.

Gillings, S., Newson, S.E., Noble, D.G. & Vickery, J.A. 2005. Winter availability of cereal stubble attracts declining farmland birds and positively influences breeding population trends. *Proc. Biol. Sci.* **272**: 733–739.

Gregory, R.D., Noble, D.G. & Custance, J. 2004. The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. *Ibis* **146**(Suppl. 2): 1–13.

Harrison, X.A. 2015. A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ.* **3**: e1114.

Hoffman, M.D. & Gelman, A. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **5**: 1593–1623.

Ives, A.R. 2015. For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods Ecol. Evol.* **6**: 828–835.

Johnson, D.E. & Fritz, L. 2014. agTrend: a Bayesian approach for estimating trends of aggregated abundance. *Methods Ecol. Evol.* **5**: 1110–1115.

Katsanevakis, K., Issaris, Y., Pouranidis, D. & Thessalou-Legaki, M. 2010. Vulnerability of marine habitats to the invasive green alga *Caulerpa racemosa var. cylindracea* within a marine protected area. *Mar. Environ. Res.* **70**: 210–218.

Kiffner, C., Lödige, C., Alings, M., Vor, T. & Rühe, F. 2011. Body-mass or sex-biased tick parasitism in roe deer (*Capreolus capreolus*)? A GAMLSS approach. *Med. Vet. Entomol.* **25**: 39–45.

Lafferty, S., Commins, P. & Walsh, J.A. 1999. *Irish Agriculture in Transition. A Census Atlas of Agriculture in the Republic of Ireland*. Maynooth: Teagasc and N.U.I.

Lee, Y. & Nelder, J.A. 2006. Double hierarchical generalized linear models (with discussion). *Appl. Stat.* **55**: 139–185.

Lindén, A. & Mäntyniemi, S. 2011. Using negative binomial distribution to model overdispersion in ecological count data. *Ecology* **92**: 1414–1421.

McCarthy, M. 2007. *Bayesian Methods for Ecology*. Cambridge: Cambridge University Press.

McCullagh, P. & Nelder, J.A. 1989. *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

McMahon, B.J., Helden, A., Anderson, A., Sheridan, H., Kinsella, A. & Purvis, G. 2010. Interactions between livestock systems and biodiversity in South-East Ireland. *Agric. Ecosyst. Environ.* **139**: 232–238.

McMahon, B.J., Anderson, A., Carnus, T., Helden, A.J., Kelly-Quinn, M., Maki, A., Sheridan, H. & Purvis, G. 2012. Different bioindicators measured at different spatial scales vary in their response to agricultural intensity. *Ecol. Indic.* **18**: 676–683.

McMahon, B.J., Carnus, T. & Whelan, J. 2013a. A comparison of winter bird communities in agricultural grassland and cereal habitats in Ireland: implications for Common Agricultural Policy reform. *Bird Study* **60**: 176–184.

McMahon, B.J., Sheridan, H., Anderson, A., Carnus, T. & Purvis, G. 2013b. Regional and farm system drivers of avian biodiversity within agriculture ecosystems. *Asp. Appl. Biol.* **121**: 203–212.

Nikoloulopoulos, A.K. & Karlis, D. 2009. Modeling multivariate count data using copulas. *Commun. Stat. Simul. Comput.* **39**: 172–187.

O'Hara, R.B. & Kotze, D.J. 2010. Do not log-transform count data. *Methods Ecol. Evol.* **1**: 118–122.

Ourens, R., Freire, J., Vilar, J. & Fernandez, L. 2014. Influence of habitat and population density on recruitment and spatial dynamics of the sea urchin *Paracentrotus lividus*: implications for harvest refugia. *ICES J. Mar. Sci.* **71**: 1064–1072.

R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B. Stat. Methodol.* **64**: 583–639.

Stan Development Team. 2015. *Stan: A C++ Library for Probability and Sampling, Version 2.8.0*. Available at: http://mc-stan.org/ (accessed 17 February 2016).

Vehtari, A., Gelman, A. & Gabry, J. 2016. *loo: Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models R package version 0.1*. Available at: https://github.com/jgabry/loo (accessed 27 February 2016).

Ver Hoef, J.M. & Boveng, P.L. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology* **88**: 2766–2772.

Watanabe, S. 2013. A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**: 867–897.

Wedderburn, R.W.M. 1974. Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61**: 439–447.

Received 15 July 2016;
revision accepted 10 December 2016.

SUPPORTING INFORMATION

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

Table S1. All species recorded in this study are presented in Table S1.