Received: 26 February 2016 | Revised: 8 August 2016 | Accepted: 12 August 2016

DOI: 10.1002/ece3.2449

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Spatial modeling of data with excessive zeros applied to reindeer pellet-group counts

Youngjo Lee[1] | Md. Moudud Alam[2] | Maengseok Noh[3] | Lars Rönnegård[2] | Anna Skarin[4]

[1]Department of Statistics, Seoul National University, Seoul, Korea

[2]School of Technology and Business Studies, Dalarna University, Falun, Sweden

[3]Department of Statistics, Pukyong National University, Busan, Korea

[4]Department of Animal Nutrition and Management, Swedish University of Agricultural Sciences, Uppsala, Sweden

Correspondence
Md. Moudud Alam, School of Technology and Business Studies, Dalarna University, Falun, Sweden.
Email: maa@du.se

Abstract

We analyze a real data set pertaining to reindeer fecal pellet-group counts obtained from a survey conducted in a forest area in northern Sweden. In the data set, over 70% of counts are zeros, and there is high spatial correlation. We use conditionally autoregressive random effects for modeling of spatial correlation in a Poisson generalized linear mixed model (GLMM), quasi-Poisson hierarchical generalized linear model (HGLM), zero-inflated Poisson (ZIP), and hurdle models. The quasi-Poisson HGLM allows for both under- and overdispersion with excessive zeros, while the ZIP and hurdle models allow only for overdispersion. In analyzing the real data set, we see that the quasi-Poisson HGLMs can perform better than the other commonly used models, for example, ordinary Poisson HGLMs, spatial ZIP, and spatial hurdle models, and that the underdispersed Poisson HGLMs with spatial correlation fit the reindeer data best. We develop R codes for fitting these models using a unified algorithm for the HGLMs. Spatial count response with an extremely high proportion of zeros, and underdispersion can be successfully modeled using the quasi-Poisson HGLM with spatial random effects.

KEYWORDS

excessive zeros, habitat preference, hierarchical generalized linear model, pellet-group count, Poisson model, spatial correlation

1 | INTRODUCTION

Fecal pellet-group counts have long been used in wildlife management to map population densities of large herbivores and their habitat selection (see, e.g., Fattorini, Ferretti, Pisani, & Sforzi, 2011; Neff, 1968; Skarin, 2008). The technique provides managers with a simple and cheap alternative to modern technologies, such as GPS collars for the surveillance of animal populations (Edge & Marcum, 1989). Although the pellet-group counts provide only a crude indication of animal habitat use rather than more precise measures of movement and habitat selection which, for example, GPS tracking can do, they still give a general idea of species distributions over different geographic areas and information about habitat use by all animals in a population using a defined area. However, an awareness of pellet decay is necessary in order to interpret the data correctly (e.g., Davis & Coulson, 2016; Skarin, 2008).

A reindeer pellet-group survey was conducted in the northern forest area of Sweden in order to assess the impact of newly established wind farms on reindeer habitat selection. From the initial survey data, collected over the 2 years 2009–2010, appearance of large numbers of 0 counts was identified as a challenge for the data analysis. This situation is not unusual in that data pertaining to spatial species counts

often contain excessive zeros (see, e.g., Agarwal, Gelfand, & Citron-Pousty, 2002; Dénes, Silveira, & Beissinger, 2015; Zuur, Saveliev, & Ieno, 2012), and requires appropriate modeling.

In the literature, zero-inflated Poisson (ZIP; Lambert, 1992), hurdle models (Cragg, 1971), and their extensions are widely suggested for modeling counts with excessive zeros (Zuur et al., 2012). Recently, Neelon, Ghosh, and Loebs (2013) used a spatial (Poisson) hurdle model to analyze hospital emergency department visits. Agarwal et al. (2002) used a ZIP model for modeling spatial species counts. Ver Hoef and Jansen (2007) used a spatiotemporal ZIP model for analyzing data on Harbor seal counts. Agarwal et al. (2002) and Neelon et al. (2013) used intrinsic conditional autoregressive (ICAR) structure, whereas Ver Hoef and Jansen (2007) used conditional autoregressive (CAR) random effects to handle spatial correlation, and all of them took a Bayesian approach to their model computations, using Markov Chain Monte Carlo (MCMC) simulations.

Even though the hurdle and ZIP models are often found to be suitable for analyzing data with excessive zero counts, these models apply only when there is overdispersion in the response variable. However, in many applications it has been found that excessive zero counts are associated with underdispersion (see, e.g., Oh, Washington, & Nam, 2006; Tin, 2008) for which ZIP and hurdle models do not make sense. Ridout and Besbeas (2004) presented two examples closely related to ecology, one pertaining to variability in birds' clutch size and another on polyspermy of eggs, which showed underdispersion. Unfortunately, any discussion of the issue of underdispersion associated with excessive counts has been missing from ecological applications. One possible reason might be that the Bayesian computational software (e.g., WinBugs; Lunn, Thomas, Best, & Spiegelhalter, 2000) restricts researchers to including only overdispersion with zero-inflated count responses. Theoretically, given that the mean is correctly specified as a function of the covariates, underdispersion can appear with count data if the underlying data are generated, for instance, from a double-Poisson (Efron, 1986), weighted Poisson (Ridout & Besbeas, 2004), Poisson mixture (e.g., generalized Poisson; Consul & Jain, 1973), or gamma (Oh et al., 2006) distribution. A failure to account for the correct type of over-/underdispersion with count data can lead to the model having a poor fit of the model (Ridout & Besbeas, 2004; Tin, 2008), very different estimates of the regression parameters (Ver Hoef & Boveng, 2007), and incorrect inferences about the model parameters (McCullagh & Nelder, 1989).

In this study, we show that high spatial correlation among counts can result in excessive zeros. Thus, to fit the reindeer pellet-group counts, we investigate whether an over-/underdispersed quasi-Poisson hierarchical generalized linear model (HGLM; Lee, Nelder, & Pawitan, 2006) with spatial correlation is suitable. The difference between a Poisson model, $y \sim \text{Poisson}(\lambda)$, and a quasi-Poisson model is that the variance of y is λ for a Poisson model, whereas the variance of y in a quasi-Poisson model includes an additional parameter ϕ such that $\text{var}(y) = \phi\lambda$ (Nelder & Pregibon, 1987). For $\phi > 1$ (or $\phi < 1$), the data are referred to as overdispersed (or underdispersed).

While a spatial hurdle model only allows for overdispersion, a spatial quasi-Poisson HGLM allows for either over- or underdispersion. The HGLM approach can confer computational advantages because the spatial quasi-Poisson HGLMs and the spatial hurdle model can both be fitted using the iterative weighted least square (IWLS) algorithm developed by Lee and Nelder (1996) for HGLMs. Furthermore, the HGLM approach enables us to compare these alternatives, via conditional Akaike's information criterion (cAIC; Lee et al., 2006) and adjusted profile likelihoods. Lee and Nelder (1996) extended the scaled deviance and its degrees of freedom for GLMs to HGLMs. Based on these, the deviance information criterion (DIC) has been proposed as a model selection criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002), while cAIC was developed as a model selection criterion in frequentist work (Vaida & Blanchard, 2005). For comparison of these two information criteria (see Section 6.5 of Lee et al., 2006).

The aim of this study is to show how a spatially correlated count response with excessive zeros can be successfully modeled using HGLMs. We achieve this by: (1) presenting the HGLMs for spatially correlated count data; (2) providing a theoretical comparison of the HGLMs with zero-inflated count data models; and (3) applying HGLM to a real data set pertaining to reindeer pellet-group counts and comparing fits of HGLMs with those of the spatial hurdle and zero-inflated models. Brief descriptions of the estimation techniques and their R (R Core Team, 2014) implementations are provided in the supplementary material. Here, it should be noted that the HGLM methodology, using IWLS, can fit the following models that include spatial correlation: Poisson generalized linear mixed model (GLMM), quasi-Poisson HGLM, and hurdle model. We use the reindeer pellet-group counts as an example showing the need for a quasi-Poisson model with spatially correlated random effects in ecological modeling. It is the only model of those investigated in this study that can accommodate both underdispersion and excess zeros.

2 | MATERIALS AND METHODS

Lee and Nelder (1996) presented HGLMs to model correlated exponential family responses by incorporating independent random effects. These models were extended to deal with correlated random effects in Lee et al. (2006). In the following, we briefly introduce a Poisson HGLM with normally distributed random effects, that is, a GLMM. Thereafter, we show how a quasi-Poisson HGLM with correlated random effects provides a way to model excess zeros jointly with over-/underdispersion. For comparison, we also present two models commonly used for data with excess zeros: the hurdle model and the zero-inflated negative-binomial model. For the different models, we also present a theory explaining how they can fit over- or underdispersion. At the end of the section, we present the example data set for reindeer pellet-group counts.

2.1 | Poisson GLMM

Consider a Poisson GLMM of counts z_i or $i = 1,2,\ldots,n$ with conditional mean $\lambda_i = E(z_i|v_i)$

$$z_i|v_i \sim \text{Poisson}(\lambda_i), \tag{1}$$

$$\eta_i = X_i\beta + v_i, \tag{2}$$

$$v_i \sim N(0,\tau), \tag{3}$$

where X_i is a row vector of covariates, and β and τ are fixed parameters.

In this model,

$$\mu_i = E(z_i) = E(E(z_i|v_i)) = \exp\left[X_i\beta + \frac{1}{2}\tau\right],$$

and

$$\text{Var}(z_i) = E(\text{Var}(z_i|v_i)) + \text{Var}(E(z_i|v_i)) = \mu_i + a\mu_i^2,$$

where $a = \exp[\tau] - 1 \geq 0$. Clearly, $\text{Var}(z_i) \geq \mu_i$ for $\tau \geq 0$, where the equality holds for $\tau = 0$. Thus, the GLMM automatically accounts for overdispersion, which resulting in more zero counts than a Poisson GLM (this issue is further elaborated in Section 2.4).

2.2 | Quasi-Poisson HGLM with spatial correlation

The spatial latent intensity approach for spatial count data was presented by Clayton and Kaldor (1987) and was subsequently modified by many others, including Cressie (1993) and Lee et al. (2006). The basic model is as follows. Given a random intensity λ_i for location i (i = 1,2,...,n), which is identified by the spatial coordinates $s(i) = (x_i, y_i)$, the conditional (count) response process z_i follows a double exponential family (Lee et al., 2006; equivalent to the extended quasi-Poisson model, Efron, 1986), that is,

$$f(z_i|v_i) = \phi^{-\frac{1}{2}} \exp\left[-\frac{\lambda_i}{\phi}\right] \frac{\exp\left[\left(\frac{1}{\phi}-1\right)z_i\right] z_i^{z_i}}{z_i!} \left(\frac{\lambda_i}{z_i}\right)^{z_i/\phi}, \tag{4}$$

$$\approx \phi^{-1} \exp\left[-\frac{\lambda_i}{\phi}\right] \frac{(\lambda_i/\phi)^{z_i/\phi}}{(z_i/\phi)!}, \tag{5}$$

where ϕ is the dispersion parameter. This model gives $E(z_i|v_i) = \lambda_i$ and $\text{Var}(z_i|v_i) = \phi\lambda_i$; ϕ = 1 gives the Poisson distribution. It allows for overdispersion when ϕ > 1 and underdispersion when ϕ < 1. Equation (5) can be obtained from equation (4) using Stirling's approximation and this was used to formulate the extended quasi-likelihood by Nelder and Pregibon (1987). Lee and Nelder (2000) showed that equations (4) and (5) give identical likelihood inferences.

Lee et al. (2006, Section 7.2) showed that the use of a quasi-Poisson-GLM can give inefficient estimate for dispersion parameter ϕ when the data are generated from a random-effect model. Zuur et al. (2012) also showed, in a simulation study, that quasi-Poisson-GLM should not be used to model overdispersion due to zero inflation. Lee et al. (2006) proposed that instead of quasi-Possion-GLM, the use of quasi-Poisson-HGLM allowing for an additional random effect v_i produces a better fit. Thus, in this study, we propose the use of quasi-Possion-HGLM for count data with excessive zeros.

Further, we model the random intensity parameter λ_i as

$$\log(\lambda_i) = X_i\beta + v_i, \tag{6}$$

where v_i is a random location effect following a certain distribution. It is generally assumed that $\mathbf{v}^T = (v_1, v_2, \ldots v_n)$ follows a multivariate normal distribution, that is, $\mathbf{v} \sim \boldsymbol{N}(0, \Sigma)$.

One popular structure of Σ for spatial covariance is $\Sigma = \tau(\mathbf{I} - \rho\mathbf{D})^{-1}$ where **I** is an identity matrix, **D** is a known symmetric matrix and Σ is positive definite giving the so called CAR structure (Besag, 1974) for **v**. However, the construction of the **D** matrix needs some careful consideration, which has been discussed elsewhere (see, e.g., Cressie, 1993; Haining, 1990; Wall, 2004).

Besides CAR (proper), other popular choices for the joint distribution of **v** includes the intrinsic CAR (or ICAR; Besag & Coperberg, 1995; Neelon et al., 2013; Lee et al., 2006) and the spatial (or simultaneous) autoregression (Ord, 1975; Wall, 2004) which gives $\mathbf{v} \sim \boldsymbol{N}(0, ((\mathbf{I} - \rho\mathbf{D})^T(\mathbf{I} - \rho\mathbf{D}))^{-1})$. All these correlation structures, CAR, ICAR, and SAR, can be fitted using a HGLM fitting algorithm (R implementations are provided in the supplementary materials).

For a CAR model (as in Section 3), following Lee et al. (2006) and Ver Hoef and Jansen (2007), we constructed $\mathbf{D} = \{d_{i,j}\}$ as

$$d_{i,j} = \begin{cases} \frac{1}{||s(i)-s(j)||} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases},$$

where $||s(i) - s(j)||$ represents the Euclidean distance between locations i and j.

2.3 | Hurdle and ZIP models for spatial zero-inflated counts

In the literature, zero-inflated Poisson (ZIP) and hurdle models, and their extensions are widely suggested for modeling counts with excessive zeros. The arguments in favor of ZIP and hurdle models, besides any background theory about the actual data generation process, are as follows. First, zero-inflated data contain more zeros than can be generated by an ordinary Poisson model for count data. Second, in the presence of zero inflation, the mean–variance relationship within an ordinary count data model breaks down. And third, the zero-inflated (and mixture) model can handle possible multiple modes (one mode at zero) in the data.

A hurdle model can be presented as follows. Given $\mathbf{v}_0 = \{v_{0,i}\}_{i=1}^n$ and $\mathbf{v}_1 = \{v_{1,i}\}_{i=1}^n$, the response follows

$$\left.\begin{array}{lll} \Pr(y_i = y|v_{0,i}, v_{1,i}, X_i) & = \mu_{0,i} & \text{if } y = 0 \\ & = (1-\mu_{0,i})\text{TP}(y,\mu_{1,i}) & \text{if } y = 1,2,\ldots, \\ g_k(\mu_{k,i}) & = X_{k,i}\beta_k + Z_{ki}\mathbf{v}_k & k = 0,1 \end{array}\right\}, \tag{7}$$

where TP is a zero-truncated Poisson probability mass function, g_k is "logit" and "log" link for k = 0 and 1, respectively, and $\mathbf{v}_k$ values are assumed to follow some sort of multivariate distribution (e.g., Gaussian CAR or SAR), which account for the spatial correlation. This specification gives

$$\mu_i^{\text{PH}} = E(y_i|v_i) = (1-\mu_{0,i})\frac{\mu_{1,i}}{1-\exp[-\mu_{1,i}]} \quad \text{and} \quad \text{Var}(y_i|v_i) = \phi_i^{\text{PH}}\mu_i^{\text{PH}},$$

where $\phi_i^{\text{PH}} = 1 + \{1 - (1-\mu_{0,i})/(1-\exp[-\mu_{1,i}])\}\mu_{1,i}$. In hurdle models, $\mu_{0,i} > \exp[-\mu_{1,i}]$ implies $\phi_i^{PH} > 1$, so that excessive zeros occur together with overdispersion.

As an alternative to the hurdle model, the ZIP and the zero-inflated negative-binomial (ZINB) models are often considered. Consider the ZINB model:

$$\left.\begin{array}{ll} P(y_i) & = p_i + (1-p_i)\exp[-\lambda_i], \quad y_i = 0 \\ P(y_i) & = (1-p_i)\frac{\exp[-\lambda_i]\lambda_i^{y_i}}{y_i!}, \quad y_i > 0 \\ \text{log it}(p_i) & = Z_i\gamma \\ \log(\lambda_i) & = X_i\beta + w_i \end{array}\right\}, \tag{8}$$

where covariates Z_i and X_i are the same as in the hurdle model and $\exp[w_i]$ follows the independent gamma distribution with $E(\exp[w_i]) = 1$. When all $v_i = 0$, it becomes the ZIP model. Since

$$\mu_i^{\text{ZIP}} = E(y_i|v_i) = p_i\lambda_i \quad \text{and} \quad \text{Var}(y_i|v_i) = \phi_i^{\text{ZIP}}\mu_i^{\text{ZIP}},$$

where $\mu_i^{\text{ZIP}} = 1 + (1-p_i)p_i\lambda_i \geq 1$, this implies that ZIP and ZINB models only allow overdispersions.

2.4 | Overdispersion due to random effects leads to higher probability of zero counts

So far, we have discussed the usefulness and limitations of ZIP and hurdle models for zero-inflated data. Now, it remains to explain how a spatial HGLM can handle excessive zero counts. We explain this issue in two steps. First, we show that overdispersion due to random effects leads to zero inflation (see Theorem 1). Then, we explain how spatial correlation can lead to even higher proportions of zeros in observed data compared with independent observations.

Theorem 1: If U_i, ($i = 1, 2, \ldots, n$) is iid Poisson-distributed with $E(U_i) = \exp[\eta_i]$, $V_i|u_i$ is also iid Poisson with $E(V_i|u_i) = \exp[\delta_i + u_i]$, $u_i \sim N(0, \sigma^2)$, and $\eta_i = \delta_i + (\sigma^2/2)$ so that $E(U_i) = E(V_i)$ but V_i is overdispersed, for $\sigma^2 > 0$ then, $\Pr(V_i = 0) > \Pr(U_i = 0)$.

Proof of Theorem 1: From the definition of marginal probability, we have

$$\begin{aligned} &\Pr(V_i = 0) = E(\Pr(V_i = 0|u_i)) = E(\exp[-\exp[\delta_i + u_i]]) \\ &= E\left(\sum_{k=0}^{\infty}\frac{1}{k!}(-\exp[\delta_i + u_i])^k\right) = \sum_{k=0}^{\infty}\frac{1}{k!}(-\exp[\delta_i])^k E(\exp[u_i]^k) \\ &\Rightarrow \Pr(V_i = 0) = \sum_{k=0}^{\infty}\frac{1}{k!}(-\exp[\delta_i])^k \exp\left(\frac{1}{2}\sigma^2k^2\right). \end{aligned} \tag{9}$$

The kth term of the summation on the right hand side of equation (9) is

$$t_k = \frac{1}{k!}(-\exp[\delta_i])^k \exp\left(\frac{1}{2}\sigma^2k^2\right). \tag{10}$$

When k is even, or 0, then $t_k \geq \frac{1}{k!}(-\exp[\delta_i])^k$ because $\exp\left(\frac{1}{2}\sigma^2k^2\right) \geq 1$ for $\sigma^2k^2 \geq 0$. Again, when k is odd, $t_k < \frac{1}{k!}(-\exp[\delta_i])^k$. But,

$$\exp[-\exp[\delta_i]] = \sum_{k=0}^{\infty}\frac{1}{k!}(-\exp[\delta_i])^k. \tag{11}$$

Comparing equations (9) and (11), we see that all the positive terms (for $k = 0, 2, 4, \ldots$) in the summation series on the right hand side of equation (9) are greater than (or equal to when $k = 0$) the corresponding terms in equation (11) and all the negative terms (for $k = 1, 3, \ldots$) in equation (9) are smaller than the corresponding terms in equation (11). Therefore, we have

$$\sum_{k=0}^{\infty}\frac{1}{k!}(-\exp[\delta_i])^k \exp\left(\frac{1}{2}\sigma^2k^2\right) > \exp[-\exp[\delta_i]] > \exp\left[-\exp\left[\delta_i + \frac{\sigma^2}{2}\right]\right].$$

But, $\Pr(U_i = 0) = \exp[-\exp[\eta_i]] = \exp[-\exp[\delta_i + \frac{\sigma^2}{2}]]$. Therefore, $\Pr(V_i = 0) > \Pr(U_i = 0)$.

From Theorem 1, we see that overdispersion due to random effects leads to a higher probability of zero counts, in other words zero inflation, compared with an ordinary Poisson GLM. To illustrate how spatial correlation can introduce an even higher proportion of zeros, let us consider two observations, Y_1 and Y_2 om a binary variable (0 represents zero counts, 1 represents nonzero), then it is straightforward to show that (proof is omitted) $\Pr(Y_1 = 0 \,\&\, Y_2 = 0|\text{Cor}(Y_1, Y_2) > 0) > \Pr(Y_1 = 0 \,\&\, Y_2 = 0|Y_1 \perp Y_2)$. In other words, due to spatial correlation, co-occurrence of zero counts can give higher proportions of zeros in some samples than expected in the case of independent observations.

A reviewer pointed out that an additional covariate could also explain excessive zeros in the data. This is true, but herein we assume that the relationships between the mean of the response and the covariates are correctly specified, and only the assumptions about the distribution should be questioned. In this respect we have in mind that in real data analysis we do not have many options about the covariates and link functions, only thing we can do is to try to improve the fit of our model by adopting different families of distributions for the response variable.

2.5 | Analysis of reindeer pellet-group counts

We analyze a real data set pertaining to reindeer fecal pellet-group counts. The data were obtained from a survey conducted on Storliden Mountain (504 m MLS; 65°13′N, 18°53′E) in northern Sweden (see Fig. 1). The size of the study area was 25 km^2, and eight windmills were built in the center of the area in 2011. The survey was conducted between 3 and 8 June in 2009 and 28 May and 1 June in 2010. Reindeer graze freely in this area from May to October except during a short period in early July when they are gathered for the marking of the calves. The survey was conducted using a point transect design (Buckland et al., 2001) and was part of a larger inventory of reindeer pellet groups over an area of 250 km^2. The distance between each transect was 300 m and the distance between each plot (red dots, in Fig. 1) on each transect was 100 m. Each plot had a size of 15 m^2 (radius = 2.18 m). The coordinates of the plots were recorded, and the center of each plot was marked with an orange wooden stick.

The pellet groups were counted using the fecal standing crop (FSC) technique in 2009 and fecal accumulation rate (FAR) in 2010 (see definitions in Skarin, 2007). A pellet group was counted for a certain plot if the center of the group was found inside the plot. Because an animal

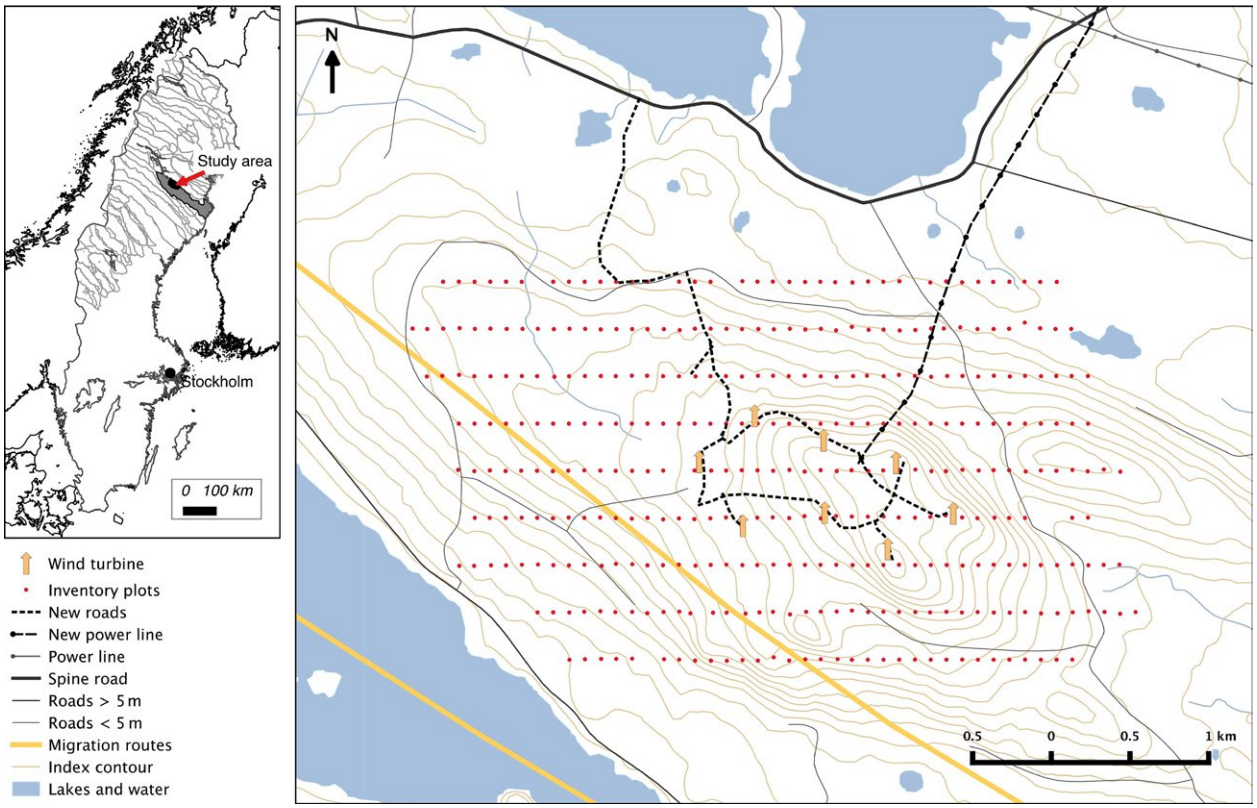

FIGURE 1 Map of the study area and the inventory plots for the pellet-group count conducted on Storliden Mountain in the Malå reindeer herding community. Location of the study area in Sweden is shown in the smaller map. © Lantmäteriet i2014/764

might move as it defecates, the pellets can spread over a large area. Therefore, a pellet group was defined by a cluster of 20 or more pellets.

Preference for habitat use by reindeer, based on the pellet-group counts, was modeled for each inventory. From the initial analysis (not reported), it was noted that 73.67% of the plots had zero counts in 2009 and 83.62% had zero counts in 2010. This indicates (possible) inappropriateness of standard count data models, for example, the Poisson GLM. Although our FSC inventory probably had a higher abundance in dry vegetation types due to the slower decay rate of the pellets (Skarin, 2007), we did not take this into account here, as the main purpose of this research was to find a method to treat the large amount of zeros in the data.

3 | RESULTS

3.1 | Spatial modeling of reindeer pellet-group counts

In order to deal with the issues of spatial correlation and excessive counts, with the pellet group data, we apply three different models for overdispersion (see Table 1). The spatial covariance structure for the normal random effects is either $\Sigma^{-1} = \frac{1}{\tau}(\mathbf{I} - \rho\mathbf{D})$ (i.e., CAR, which includes Poisson-normal HGLM as a special case for $\rho = 0$) or $\Sigma^{-1} = \sigma^2((\mathbf{I} - \rho\mathbf{D})^T(\mathbf{I} - \rho\mathbf{D}))^{-1}$. We also fit a Poisson GLM, as would be done in regression kriging (Bivand, Pebesma, & Gomez-Rubio, 2008). A detailed list of the fitted models is presented in Table 1.

As GLMs are uniquely specified by the mean (in our case, λ_i) and the variance function, $V(\lambda_i)$, we use only these parameters and functions to specify the models. All the spatial models presented in Table 1 can be fitted using the HGLM algorithm (an R implementation is available in the supplementary material).

For the 2009 data, we started with a large (full) model containing 13 covariates: the (log-) distance from the power grid, slope of the location, a ruggedness index, elevation, forest age structure, dummies (1/0) for clear-cuts, young forest, coniferous forest, broad-leaved forest, flat area, southeast slope, northwest slope, and northeast slope. To avoid possible multicollinearity, we did not include (log) distance to nearby big infrastructure (which had a correlation of 0.8 with log-distance to power grid, while the correlation between any other pairs of variables did not exceed 0.6, in absolute value) in the model. For the 2010 data, we excluded the dummy variable for broad-leaved forest from the full model. In the 2010 data, broad-leaved forests had only pellet-group counts of zero; this indicates that special care is necessary to tackle the exploding tendency of the ML estimate, if it exists, for the relevant parameter (Feinberg & Rinaldo, 2007). However, we found that the MLEs of the other parameters remained similar after the broad-leaved forest was dropped from the model.

Table 2 reveals that the quasi-Poisson-normal HGLM with $\phi < 1$ and the CAR (QCAR) specification (Model V) fits the data best, as it has the lowest cAIC. The models with the SAR covariance structure are not able to produce a better fit than the QCAR model. We do not, therefore, report those results in this paper.

TABLE 1 Specifications of the fitted models

Model	Description	Mean ($\eta_i = \log(\lambda_i)$)	Variance function ($V(\lambda_i)$)	Random effects and their distributions
I	Poisson GLM	$X_i\beta$	λ_i	No
II	Poisson HGLM	$X_i\beta + v_i$	λ_i	$\mathbf{v}^T = (v_1, v_2, \ldots, v_n)$, $v_i \sim N(0, \tau)$
III	Poisson HGLM with CAR	Model II	λ_i	$\mathbf{v} \sim N(0, \Sigma)$, $\Sigma^{-1} = (1/\tau)(\mathbf{I} - \rho\mathbf{D})$
IV	Negative-binomial HGLM with CAR	$X_i\beta + v_i + w_i$	λ_i	$\exp[w_i] \sim \text{Gamma}(\alpha, 1/\alpha)$, $\mathbf{v}$ is as in Model III
V	Quasi-Poisson HGLM with CAR	Model II	$\phi\lambda_i$	Same as Model III

TABLE 2 cAIC with full set of covariates

	cAIC[a]	
Model	2009 data	2010 data
Model I	628.71	432.13
Model II	587.93	402.84
Model III	586.55	406.15
Model IV	542.10	326.18
Model V	528.87	206.77

[a]cAIC iss the AIC for model I.

Starting with the best-fitted full model, we gradually delete covariates one at a time from the model on the basis of the absolute *t*-value (covariate with the lowest absolute *t*-value deleted first as commonly suggested in the literature, see, e.g., McCullagh & Nelder, 1989; Ch. 3.9) until we obtain the final model, having all the fixed-effect parameters significant at the 5% level (both in the Wald and likelihood-ratio tests). The estimated parameters and their standard errors for the final models for the 2009 and 2010 data are presented in Table 3.

From the results (Table 3), we see that distance from power grid was the most influential factor (both in 2009 and 2010), and it was also statistically significant (*p*-value < .001 for both GLM and QCAR). Its positive coefficient estimate reveals that the pellet-group counts were higher at locations farther away from power lines.

A plot of the observed responses against the fitted values (for the 2009 FSC count) is given in Fig. 2. The same plots for the 2010 FAR count data reveal the same overall pattern, so these plots are not shown. From Fig. 2, we see that the fit of the model gradually improves as the spatial dependence structures we incorporate become more reasonable. This indicates the advantage of joint modeling of the mean and the covariance. By comparing the plots for the four models in Fig. 2, we see that QCAR (lower right in Fig. 2; which is also the best fit model in terms of cAIC, see Table 2) not only improves the mean prediction but also reduces the predicted mean square error (PMSE; see Table 4).

Comparing the four plots of the observed counts with the in-sample fitted values in Fig. 2, we see that the simple Poisson-normal HGLM (upper left plot in Fig. 2) was not able to model excessive zero counts, adequately. Harrison (2014) also showed, in a simulation study, that this Possion-normal HGLM failed to reduce bias in zero-inflated data. By modeling spatial correlation (Poisson-normal HGLM with CAR), we get a better fit compared with the Poisson-normal HGLM with independent random effects. Figure 2 also shows that the excessive overdispersion that results from using a negative-binomial-normal HGLM with CAR does not improve the prediction. Finally, QCAR gives the best fit. The same findings hold for both the 2009 and 2010 data. Here, $\hat{\phi} = 0.737 < 1$ for 2009 and $\hat{\phi} = 0.476 < 1$ for 2010.

3.2 | Comparison with the hurdle model

The hurdle model is frequently used for analyzing count response with excess zeros. Therefore, we also analyze the reindeer pellet-group counts using such a model (eq. 7). We assume $\mathbf{v}_k \sim N(0, \Sigma_k)$ where Σ_1 has a CAR specification as in Model III (see Table 1). Because no R package module is able to fit Model (7) with CAR random effects in a non-Bayesian manner, we developed our own R codes to carry out the model computation using a hierarchical likelihood (h-likelihood) approach (a brief description of the algorithm, and the R program are available in the supplementary material).

For the binary part ($\mu_{0,i}$ in eq. 7), we use four covariates (northwest slope, southeast slope, elevation, and log-distance to power grid) for the 2009 data and six covariates (southeast slope, Young forest, clear-cuts, forest age structure, elevation, and log-distance to power grid) for the 2010 data. These variables were selected on the basis of separate binomial models for Pr(Count > 0) and because they have the lowest cAIC values. For the truncated Poisson model in equation (7), we use the same set of covariates as in Table 3. Because the truncated Poisson part of the hurdle model does not provide a direct estimate of the random effects for the locations with observed 0 counts, a direct computation of the fitted values for the hurdle model is not straightforward (a kriging approach as presented in Section 3.4 could be an option though). Therefore, we report the fit of the binomial and the truncated Poisson part of the hurdle models separately in Fig. 3. The top panel of Fig. 3 shows the fit of the binomial (left) and truncated Poisson (right) parts of the hurdle model for the 2009 data. The bottom panel of Fig. 3 shows the same for the 2010 data.

Comparing Figs 2 and 3, we see that the HGLM with QCAR (Model V) provided a better fit than the hurdle model. Although the truncated Poisson part of the hurdle model did a reasonably good job, the failure in the binary part downgraded the overall prediction. We tried, using all the available variables, to improve the performance of the binary part of the model, but we failed. We could not fit a spatial correlation in the binary part because standard R packages, for example, lme4 (Bates, Maechler, Bolker, & Walker, 2015) and hglm (Alam, Ronnegard,

TABLE 3 Estimated model parameters and fit statistics for Model I and Model V (final)

	For 2009 FSC counts		For 2010 FAR counts	
Parameters	**Model I**	**Model V**	**Model I**	**Model V**
Intercept	−18.916 (3.782)	−18.171 (4.491)	−12.21 (4.562)	−10.979 (5.163)
Northwest slopes	−0.489 (0.304)	−0.656 (0.364)		
Southeast slope			0.696 (0.316)	0.988 (0.410)
Elevation	0.007 (0.002)	0.007 (0.003)	−0.005 (0.003)	−0.005 (0.004)
Distance to power lines	1.897 (0.426)	1.728 (0.519)	1.569 (0.499)	1.33 (0.595)
Clear-cuts			0.607 (0.346)	0.567 (0.500)
τ		1.324 (0.270)		2.327 (0.393)
ϕ		0.737 (0.082)		0.476 (0.036)
ρ		3.038 (0.750)		1.903 (4.416)
cAIC[a]	618.4	526.25	418.78	221.44

Values in parentheses represent standard error.
[a]cAIC is AIC for GLM.

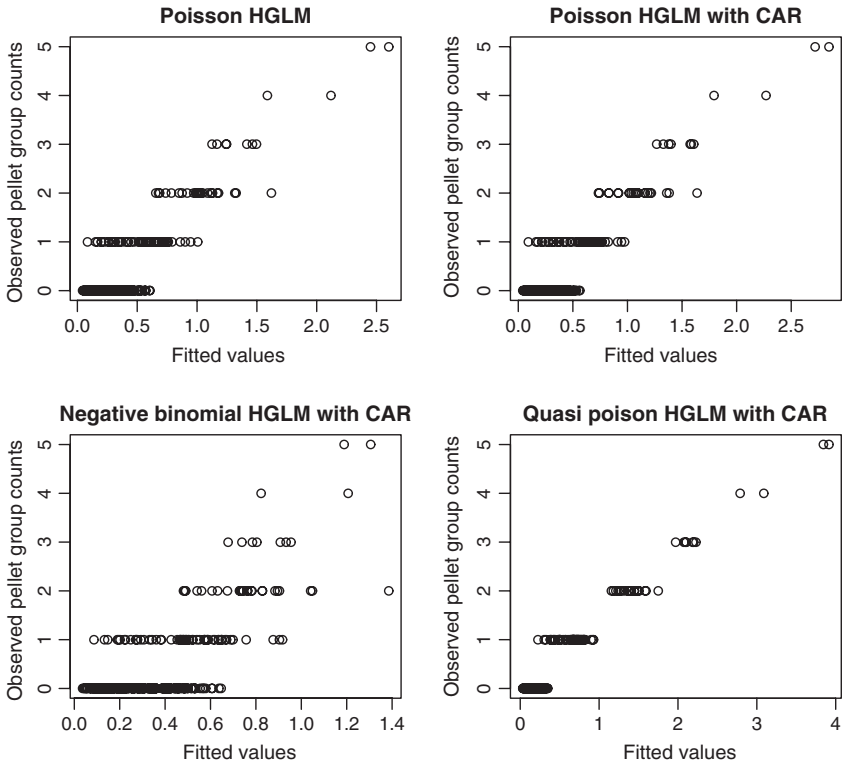

FIGURE 2 Plots of observed versus fitted values for different models (with the full set of covariates) for the 2009 FSC count

& Shen, 2015), cannot fit a binary GLMM with one observation per subject. If repeated observations from each plot were available, we could try to apply bivariate random effects (for binary part and Poisson part), as used in Neelon et al. (2013), but with the current data set we cannot improve on our current approach.

3.3 | Comparison with a zero-inflated Poisson model

Figure 4 shows that the HGLM with QCAR (Model V) provided a better fit than Poisson GLM, ZIP, and ZINB models. The hurdle model allows overdispersion only for zero-deflation cases. However, this data set exhibits underdispersion but with excessive zeros. This means that ZIP, ZINP, and hurdle models are not appropriate to apply to this data set. The quasi-Poisson HGLM allows for underdispersion (ϕ < 1) with excessive zeros.

3.4 | Prediction by the models

To evaluate the performance of predictions from various models, the whole data set is divided randomly into two parts: 70% as the data form training set and the remaining 30% form the test set; this division is repeated for 100 times. After fitting models I–V to each training set,

TABLE 4 Average PMSE of various models, based on 100 random test data sets

Model	PMSE for 2009 data	PMSE for 2010 data
Model I	0.665	0.351
Model II	0.663	0.349
Model III	0.272	0.201
Model IV	0.142	0.154
Model V	0.123	0.115
Hurdle	0.606	0.380[a]
ZIP	0.620	0.355
ZINB	0.621	0.355

[a]Calculated after ignoring four abnormal (>1,000) PMSEs.

the PMSE is computed for the rest of the data (the test set) using the following formula

$$\mathrm{PMSE} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} (y_i - \hat{\lambda}_i)^2, \quad (12)$$

where n_{test} is the sample size of the test set, y_i is the response of the test set and $\hat{\lambda}$ is the estimator of λ_i using the training set. For models I–II without a spatial correlation, $\lambda_i = \exp[X_i\hat{\beta}]$ where $\hat{\beta}$ is the estimated β from the training set. For models III–V with a spatial correlation, $\lambda_i = \exp[X_i\hat{\beta} + \hat{v}_i]$ where $\hat{v}_i$ is the predicted value of v_i. We compute $\hat{v}_i = \widehat{\mathrm{cov}}(v_i, v_i^{\mathrm{Train}})\widehat{\mathrm{cov}}(v_i^{\mathrm{Train}})^{-1}\hat{v}_i^{\mathrm{Train}}$ where $\hat{v}_i^{\mathrm{Train}}$ is the random location effect of the training set. Here, $\widehat{\mathrm{cov}}(v_i^{\mathrm{Train}})^{-1}$ and $\hat{v}_i^{\mathrm{Train}}$ are estimated using the training set, and $\widehat{\mathrm{cov}}(v_i, v_i^{\mathrm{Train}})$ is $\mathrm{cov}(v_i, v_i^{\mathrm{Train}})$ after replacing the parameters involved with their estimates from the training set.

Table 4 shows the average PMSE of 100 random selections of the training and test sets. HGLM with QCAR (Model V) with underdispersion and spatial correlation is, overall, the best-fitting model giving the lowest cAIC (also known as DIC, see Lee et al., 2006, Ch. 6.5; see also the supplementary materials) and PMSE, together. Thus, the model with underdispersion gives better predictions than overdispersed models with zero inflation.

4 | DISCUSSION

In this paper, we introduce a quasi-Poisson HGLM with a spatial correlation to fit reindeer pellet-group counts, and we show that a Poisson GLM, by ignoring spatial correlation, can lead to a poor model fit. Such a simplified model produces poor-quality residuals (due to the lack of fit), which lead to incorrect conclusions being drawn about the spatial correlation. Consequently, the regression kriging prediction based upon those residuals, which is often suggested in the literature (see, e.g., Bivand et al., 2008; Cressie, 1993; Gribko, Hohn, & Ford, 1999), may result in poor spatial prediction.

From the results of the fitted quasi-Poisson HGLM (see Table 3), we conclude that several environmental variables, for example, slope, elevation, and vegetation type at the location, as well as human development activities, for example, power lines, are significant factors explaining reindeer habitat preference.

In the literature, hurdle and ZIP models are widely used for analyzing count responses with excessive zeros. However, hurdle and ZIP models do not allow for underdispersion with excessive zeros. In practice, such data sets often exist, for example, the incidence rate of

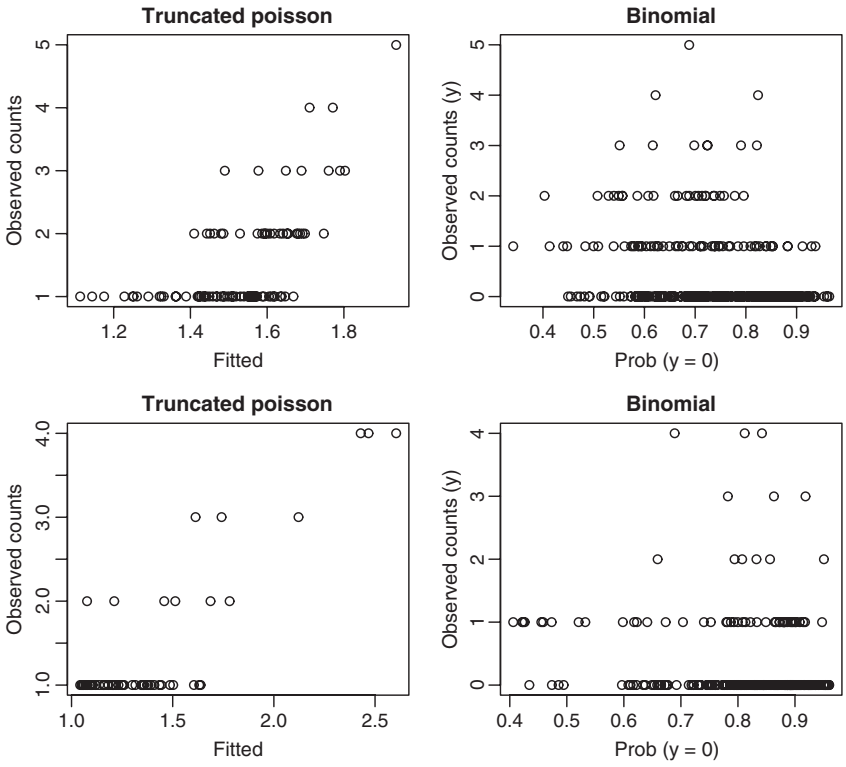

FIGURE 3 Plots of observed versus fitted values for the binomial and truncated Poisson parts of the hurdle models for the FSC (Year 2009, top panel) and FAR (Year 2010, bottom panel) counts

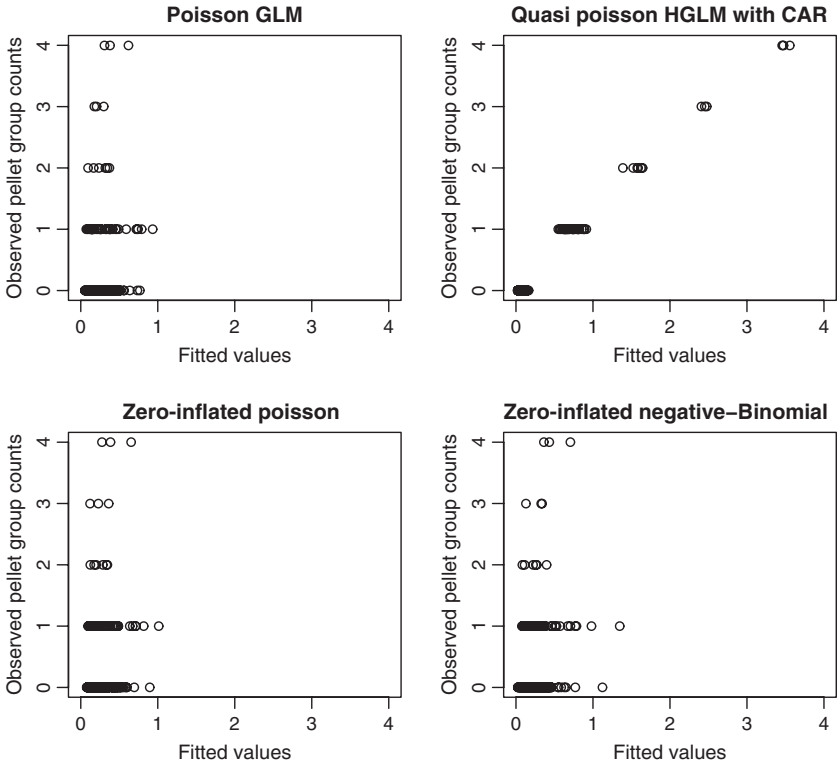

FIGURE 4 Plots of fitted versus observed values for Poisson GLM, QCAR, ZIP, and ZINB for the 2010 FAR counts

hospitalization (Tin, 2008), and accident rates when accidents are very rare events (Oh et al., 2006), where excess zero counts appear along with underdispersion.

For a real data set pertaining to pellet-group counts, we fit models with zero inflation, ZIP and ZINB. However, they give poor predictions (Table 4) and have poor fitted values (Fig. 4). Spatial-correlation-only models (III and IV) improve both prediction and fitted values. However, hurdle, a model with both zero inflation and spatial correlation, is worse than a spatial-correlation-only model. We show that Poisson HGLM with spatial correlation and underdispersion, namely Model V, provides the best predictions and fitted values.

With the reindeer pellet-group counts, a quasi-Poisson HGLM with CAR allows for underdispersion ($\phi < 1$) with excessive zeros for both 2009 and 2010 data. Thus, quasi-Poisson HGLM with CAR random effects provides a better fit with the data than the hurdle model with similar linear covariates and correlation structures. Our results, however, do not imply that Poisson HGLM can be safely used to analyze data that are generated by a true zero-inflated Poisson or a hurdle model (see contrasting example in Zuur et al., 2012). If the underlying subject matter theory leads to a ZIP or a hurdle model, then that model should be applied. However, the results show that we cannot reject a HGLM in favor of a ZIP or a hurdle model only because the data contain a high proportion of zeros; overdispersion, high correlation, and a covariate may well be able to explain the excessive zeros.

It would be interesting in future work to extend hurdle models to allow for underdispersion with excessive zeros by adopting some sort of weighted Poisson distribution (Ridout & Besbeas, 2004; and references cited therein), generalized Poisson, or gamma distribution (Oh et al., 2006) for the positive response part. However, computation of those models, especially when there is spatial dependence, and the interpretation of the model parameters would be challenging tasks. Spatial hurdle models are commonly fitted using Bayesian MCMC techniques (Zuur et al., 2012), which are computationally too intensive. If there is no special reason for using a Bayesian approach (such as priors originating from a theoretical justification), one can use an HGLM model computed using the h-likelihood method that provides a deterministic algorithm (R code is provided in the supplementary material with this paper) and is faster than conventional MCMC methods.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2011-0030810). This research was also supported in part by the Brain Research Program through the NRF funded by the Ministry of Science, ICT and Future Planning (2014M3C7A1062896), the Swedish Agency of Energy, and the research programme "Vindval" (research grant no. 31940-1).

CONFLICT OF INTEREST

None declared.

DATA ACCESSIBILITY

The reindeer pellet-group survey data set used in this article is available in the supplementary material.

REFERENCES

Agarwal, D. K., Gelfand, A. E., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, *9*, 341–355.

Alam, M., Rönnegård, L., & Shen, X. (2015). Fitting conditional and simultaneous autoregressive spatial models in hglm. *The R Journal*, *7*, 5–18.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, *36*, 192–236.

Besag, J., & Coperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*, 733–746.

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2008). *Applied spatial data analysis with R*. New York, NY: Springer.

Buckland, S., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to distance sampling-estimating abundance of biological populations*. New York, NY: Oxford University Press.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimation of age-standardized relative risk for use in disease mapping. *Biometrics*, *43*, 671–681.

Consul, P. C., & Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, *15*, 791–799.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*, 829–844.

Cressie, N. A. C. (1993). *Statistics for spatial data (revised)*. New York, NY: Wiley.

Davis, N. E., & Coulson, G. (2016). Habitat-specific and season-specific faecal pellet decay rates for five mammalian herbivores in south-eastern Australia. *Australian Mammalogy*, *38*(1), 105–116.

Dénes, F. V., Silveira, L. F., & Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: Accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, *6*, 543–556.

Edge, W. D., & Marcum, C. L. (1989). Determining elk distribution with pellet-group and telemetry techniques. *The Journal of Wildlife Management*, *53*, 621–624.

Efron, B. (1986). Double exponential families and their use in generalized linear models. *Journal of the American Statistical Association*, *81*, 709–721.

Fattorini, L., Ferretti, F., Pisani, C., & Sforzi, A. (2011). Two-stage estimation of ungulate abundance in Mediterranean areas using pellet group count. *Environmental and Ecological Statistics*, *18*, 291–314.

Feinberg, S. E., & Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and categorical data analysis. *Journal of Statistical Planning and Inference*, *137*, 3430–3445.

Gribko, L. S., Hohn, M. E., & Ford, W. F. (1999). White-tailed deer impact on forest regeneration: Modeling landscape-level deer activity patterns. In Stringer, W. Jeffrey, L. Loftis, David, (Ed.) *Proceedings, 12th central hardwood forest conference*. (pp. 178–185). Asheville, NC: Department of Agriculture, Forest Service, Southern Research Station.

Haining, R. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge, UK: Cambridge University Press.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *3*, 1–14.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, *58*, 619–656.

Lee, Y., & Nelder, J. A. (2000). Two ways of modeling overdispersion in non-normal data. *Journal of the Royal Statistical Society, Series C*, *49*, 591–598.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with random effects: Unified analysis via h-likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

Lunn, D. J., Thomas, J. A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London, UK: Chapman & Hall.

Neelon, B., Ghosh, P., & Loebs, F. P. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society, Series A*, *176*, 389–413.

Neff, D. J. (1968). The pellet-group count technique for big game trend, census, and distribution: A review. *The Journal of Wildlife Management*, *32*, 597–614.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*, 221–232.

Oh, J., Washington, S. P., & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis and Prevention*, *38*, 346–356.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*, 120–126.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Ridout, M. R., & Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, *4*, 77–89.

Skarin, A. (2008). Decay rate of reindeer pellet-groups. *Rangifer*, *28*, 47–52.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, *64*, 583–640.

Tin, A. (2008). *Modeling zero-inflated count data with underdispersion and overdispersion*. SAS Global Forum 2008: Statistics and Data Analysis. Retrieved from http://www2.sas.com/proceedings/forum2008/372-2008.pdf

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*, 351–370.

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

Ver Hoef, J. M., & Jansen, J. K. (2007). Space-time zero-inflated count models of Harbor seals. *Environmetrics*, *18*, 697–712.

Wall, M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*, 311–324.

Zuur, A. F., Saveliev, A. A., & Ieno, E. N. (2012). *Zero inflated models and generalized linear mixed models with R*. Newburgh, UK: Highland Statistics Ltd.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Lee, Y., Alam, M. M., Noh, M., Rönnegård, L. and Skarin, A. (2016), Spatial modeling of data with excessive zeros applied to reindeer pellet-group counts. Ecology and Evolution, 6: 7047–7056. doi: 10.1002/ece3.2449