Working papers in transport, tourism, information technology and microdata analysis

Analyzing spatially correlated counts with excessive zeros: a case of modeling the changes of reindeer distribution

Youngjo Lee
Md. Moudud Alam
Maengseok Noh
Lars Rönnegård
Anna Skarin
Editor: Hasan Fleyeh

Nr: 2013:06

Working papers in transport, tourism, information technology and microdata analysis
ISSN: 1650-5581

Analyzing spatially correlated counts with excessive zeros: a case of modeling the changes of reindeer distribution

Youngjo Lee
Department of Statistics, Seoul National University, Seoul, Republic of Korea. e-mail: *youngjo@snu.ac.kr*
Md. Moudud Alam[1]
Dalarna University, School of Technology and Business Studies, Borlänge, Sweden. e-mail: *maa@du.se*
Maengseok Noh
Department of Statistics, Pukyong National University, Busan, Republic of Korea. e-mail: *msnoh@pknu.ac.kr*
Lars Rönnegård
Dalarna University, School of Technology and Business Studies/Statistics, Borlänge, Sweden. e-mail: *lrn@du.se*
Anna Skarin
Department of Animal Nutrition and Management, Swedish University of Agricultural Sciences, Uppsala, Sweden. e-mail: *anna.skarin@slu.se*

Abstract

Spatial dependency is a common issue in the modeling of ecological data, especially in the surveying of animal distributions. In this paper, we argue that when we make such an inference we should account for both environmental factors and spatial correlation. We show how this can be done by using hierarchical generalized linear models (HGLMs), which allow us to model wide classes of spatial dependencies. In a real data set of reindeer fecal pellet-group count at sample locations in a northern Swedish forest, we found that over 70% of counts were zeros. Analyzing this data set we show that the proposed HGLM-based models can perform better than the other commonly used models, e.g. ordinary Poisson model and spatial hurdle model, in modeling spatially correlated count data with excessive zeros.

keywords: Excessive spatial correlation, HGLM, pellet-group count, reindeer habitat preference

[1] Corresponding Author. Author's contribution to this work was initiated during his post doctoral visit at Seoul National University.

1 Introduction

Spatial dependency is a common issue in the modeling of ecological data, especially in the surveying of animal or other species distributions (Dolan et al. , 2000). Fecal pellet-group counts have long been used in wildlife management to map population densities of large herbivores and their habitat selection (e.g. Fattorini et al. , 2011; Neff , 1968; Skarin , 2007). The technique provides management with a simple and cheap alternative to modern technologies such as GPS collars for the surveillance of animal populations (Edge and Marcum , 1989).

To handle possible spatial correlation within fecal pellet-group data, earlier efforts have been made via the geostatistical technique of ordinary kriging of samples (Gribko et al. , 1999). An extension of ordinary kriging would be the so called regression krigging (Cressie , 1993) which enables to include dependence of environmental variable through a suitable regression model. In application, regression krigging is implemented in a two step procedure where the regression model is fitted in the first step by ignoring the spatial correlation and then, in the second step, the residuals from the regression model is used to estimate any spatial dependence (see e.g. Bivand et al. , 2008). Here, we propose new insights into the possibility of analyzing the spatial distribution of pellet groups both environmentally and spatially using hierarchical generalized linear models (HGLM; Lee and Nelder , 1996) for joint modeling of the spatial trend and the spatial covariance. The advantage of being able to model fecal pellet groups is that it allows one to compare differences in spatial distribution over the years if sampling is repeated in the same area.

We demonstrate how different parsimonious, yet intuitively appealing, spatial correlation structures can be modeled with HGLMs. Further, the techniques of model selection for spatial models for both the mean part and the correlation structure are demonstrated. These models lead to various forms of covariance structures. By analyzing a real dataset on Reindeer pellet group counts, we show that HGLMs can be a useful tool for modeling spatially correlated count data with excessive zeros.

In literature, zero inflated Poisoon (ZIP; Lambert , 1992) and hurdle models (Cragg , 1971) are widely suggested for modeling counts with excessive zeros. Recently, spatial Poisson hurdle models are used for analyzing count data with excess zeros in, among others, Neelon et al. (2012) and Agarwal et al. (2002). Therefore, we also analyze the real data sets with spatial Poisson hurdle model. The results show that HGLM model outperforms spatial hurdle model in terms of the fit of the model.

The aim of this paper is to show how spatially correlated count response with excessive zeros can be successfully modeled using HGLMs. We will demonstrate this by i) presenting the HGLMs for spatially correlated count data (Section 2), ii)presenting the estimation and techniques of model comparison (Section 3) and iii)applying them to a real data set on pellet group counts and comparing fits of HGLM models with that of the spatial hurdle model (Section 4).

2 HGLM models for spatial data

The HGLMs to model correlated exponential family responses, by using independent random effects, were presented by Lee and Nelder (1996). These models are extended to deal with correlated random effects in Lee et al. (2006). In the following, we show that the HGLMs with correlated random effects provides a way to jointly model spatial trend and correlation parsimoniously, without compromising the intuitive appeal of the models.

The spatial latent parameter approach for a spatial count data model was presented by Clayton and Kaldor (1987) and was further discussed and 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 $i\,(i = 1, 2 \ldots, n)$, which is identified by the spatial coordinates $s_i = (x_i, y_i)$, the conditional (count) response process z_i follows a Poisson or double exponential family (Efron , 1986, equivalent to the extended quasi Poisson model: Lee et al. 2006) i.e., the conditional density for the Poisson exponential family is

$$f\left(z_i|\lambda_i\right) = \frac{\exp\left[-\lambda_i\right]\lambda_i^{z_i}}{z_i!} \tag{1}$$

and for a double Poisson exponential family

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

$$\approx \quad \phi^{-1} \exp\left[-\frac{\lambda_i}{\phi}\right] \frac{\left(\frac{\lambda_i}{\phi}\right)^{z_i/\phi}}{\left(\frac{z_i}{\phi}\right)!} \tag{3}$$

where ϕ is the dispersion parameter; $\phi = 1$ gives the Poisson distribution (1). Equation (3) is the extended quasi-likelihood from Nelder and Pregibon (1987), which can be obtained from equation (2) using Stirling's approximation. Therefore, both equations should give similar likelihood inferences (see Lee et al. , 2006).

Further, we model the random intensity parameter λ_i as

$$g(\lambda_i) = X_i\beta + h(u_i) \tag{4}$$

where g and h are (known) monotone link functions and u_i is a random location effect, which follows a certain distribution. For a Poisson model, $g(\lambda_i) = \log(\lambda_i)$ gives the canonical link (Lee et al. , 2006). We often assume $\mathbf{u}^T = (u_1, u_2, \ldots, u_n)$ follows a multivariate normal distribution, i.e., $\mathbf{u} \sim N(\mathbf{0}, \Sigma)$, and h to be the identity link, i.e. $h(u) = u$. Though the normality of $\mathbf{u}$ is not required, we demonstrate how the different types of spatial dependences in z can be parsimoniously modeled by imposing a certain structure on the Σ matrix under the normality assumption.

2.1 Modeling spatial correlation

In principle, Σ can be any symmetric and positive definite matrix for the multivariate normal $h(u) = u$. However, an unstructured Σ is not an attractive choice, because we would need to estimate $n(n+1)/2$ free parameters. In this section, we present the method for modeling different spatial correlation structures in terms of a few parameters in Σ.

Assume that the latent intensity for location i is composed of a fixed part $X_i\beta$ and an independent random (latent or unobservable location-specific advantage or disadvantage) part $\varepsilon_i \sim N(0, \sigma^2)$. If the locations are close to each other then any (latent) advantage (or disadvantage) of location i may partially be enjoyed by location j $(j = 1, 2, \ldots, n)$, where the exact amount of transmission of the (latent) advantage is proportional to the distance between the two sites i and j. This leads to the following construction:

$$u_i = \varepsilon_i + \sum_{i \neq j = 1}^{n} k(|s_i - s_j|)\, \varepsilon_j \tag{5}$$

where $|s_i - s_j|$ denotes distance (with respect to some distance measure) between location i and j having spatial coordinates s_i and s_j, and $k(|s_i - s_j|)$ is any symmetric function of the distances such that $k(|s_i - s_j|) = k(|s_j - s_i|)$. In the application it makes sense to use

$$k(|s_i - s_j|) = \begin{cases} \frac{1}{||s_i - s_j||^c} & \text{if } ||s_i - s_j|| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $||s_i - s_j||$ denotes the Euclidean distance and c is a constant. In principle, c can be estimated from the data however in this paper we assume c = 1(known). In doing so, we are essentially assuming that each location receives a random effect of its own plus a weighted sum of the independent random effects of all the locations, with the weights being the inverse of the distances between the current location and the other locations.

Equation (5) gives the following covariance matrix of the random effects:

$$\Sigma = \sigma^2 (I + D)(I + D)^T \tag{7}$$

where I is an $n \times n$ identity matrix and the $(i, j)^{th}$ element of the matrix D is $k(|s_i - s_j|)$.

Although the autoregressive model (7) has an intuitive explanation, it has several drawbacks, e.g., it does not contain the independent random effects model as a special case unless $D = 0$, and it is too simplistic in that the correlation between the locations are fully determined by the known distance matrix D.

A second proposal for the covariance structure of the random effects comes from the motivation of the Markov random field (MRF) model. A Markov random field is a stochastic process whose current state, given a few neighboring states, does not depend on the other states (Kindermann and Snell , 1980). Because the inverse of the covariance matrix (also known as the precision matrix) of a multivariate normal distribution determines the partial correlations, it is easy to specify a MRF model by using the structured precision matrix. Denoting the $(i,j)^{th}$ element of Σ^{-1} as $\Sigma^{-1}_{i,j}$ we notice that $\Sigma^{-1}_{i,j} = 0$ implies that the partial correlation between u_i and u_j is 0, given the realizations of all other random effects. Hence, by restricting certain elements of Σ^{-1} to zero, we can produce a MRF specification for $\mathbf{u}$.

For spatial data, a MRF model may assume a precision matrix

$$\Sigma^{-1} = \frac{1}{\tau}\left(I - A\left(\rho\right)\right) \tag{8}$$

where τ and ρ are parameters, and the matrix $A\left(\rho\right)$ may be structured as $A\left(\rho\right)) = \rho N$, where N is the so-called neighborhood matrix, whose diagonal elements are all 0 and the $(i,j)^{th}$ off-diagonal element is 1 if location i and j are neighbors, and 0 otherwise (Clayton and Kaldor , 1987, see e.g.,). Note that MRF models with $\rho = 0$ reduce to classical HGLMs with independent random effects.

The precision matrix in (8) makes sense for spatial data on a lattice (Kindermann and Snell , 1980; Cressie , 1993). However, the neighborhood alone may not be enough to determine the correlation among the locations (see e.g., Leichstein et al. , 2002) and the neighbors are not always well-defined (as happens when the sample locations do not share common physical borders between them). In such cases, an immediate extension of $A\left(\rho\right)$ can be given as $A\left(\rho\right) = \rho N \odot D$, where D is the matrix in the covariance model (6) and $\odot$ represents element-by-element multiplication. Another extension can be given as $A\left(\rho\right) = \rho D$. In the latter case we are assuming that the partial correlation between two locations decays with increasing distance, but may not vanish. Consequently, the last specification does not belong to the MRF model class, but it gives conditional autoregressive specification (CAR; Clayton and Kaldor , 1987; Xiaoping et al. , 2007) for the latent Poisson intensity, λ_i.

2.2 Further extension of spatial models

Though the Poisson-normal HGLMs are intuitively appealing, HGLMs allow models with more general variance-covariance structures than do Poisson-normal spatial models. We show how such models explain excessive zeros in count responses.

Let us start with a Poisson-normal HGLM with $E\left(z_i|u_i\right) = \lambda_i$, $\log\left(\lambda_i\right) = \eta_i = X_i\beta + u_i$, $u_i \sim N\left(0, \sigma^2\right)$ and $Var\left(z_i|u_i\right) = \lambda_i$. This gives

$$E\left(z_i\right) = E_{u_i}\left(E\left(z_i|u_i\right)\right) = \exp\left[X_i\beta + \frac{1}{2}\sigma^2_u\right] = \mu_i(\text{say})$$

and

$$\begin{aligned} Var\left(z_i\right) &= E\left(Var\left(z_i|u_i\right)\right) + Var\left(E\left(z_i|u_i\right)\right) \\ &= \exp\left[X_i\beta + \frac{1}{2}\sigma^2_u\right] + \exp\left[2X_i\beta + \sigma^2_u\right]\left(\exp\left[\sigma^2_u\right] - 1\right) \\ &= \exp\left[X_i\beta + \frac{1}{2}\sigma^2_u\right] + \left(1 + \exp\left[X_i\beta + \frac{1}{2}\sigma^2_u\right]\left(\exp\left[\sigma^2_u\right] - 1\right)\right) \\ &= \mu_i + a_i\mu_i^2 \end{aligned}$$

where $a_i = \exp\left[\sigma^2_u\right] - 1$. Clearly, $Var\left(z_i\right) \geq \mu_i$ for $\sigma^2_u \geq 0$, where the equality holds for $\sigma^2_u = 0$. Thus, the HGLM automatically accounts for overdispersion. However, we found that this Poisson-normal HGLM may not be sufficient to explain excessive zero counts, such as the ones that appeared in the reindeer pellet-group count data (see Section 4).

To account for extra-Poisson overdispersion in count data, Lee and Nelder (2001) proposed two alternative approaches: a quasi-Poisson model (McCullagh and Nelder , 1989) and a Poisson-gamma model; the latter leading to a negative-binomial marginal model. In the following lines, we briefly discuss how these approaches incorporate different variance-covariance structures with HGLMs (Molenberghs et al. , 2007, see also).

Suppose that $z_i|u_i$ follows the extended quasi-Poisson model (Nelder and Pregibon , 1987), equivalent to the double exponential family of (Efron , 1986). Then, we have $Var(z_i|u_i) = \phi\lambda_i$ which gives an extra dispersion parameter ϕ in the conditional model of $z_i|u_i$. Alternatively, we may assume that $z_i|X_i, u_i, v_i$

Table 1: Differences in modeled overdispersion

Conditional response	Spatial random effect	Extra iid random effect	$E(z_i)$	$Var(z_i)$
Poisson	$N\left(0,\sigma_u^2\right)$	-	μ_i	$\mu_i + a_i\mu_i^2$
Quasi-Poisson	$N\left(0,\sigma_u^2\right)$	-	μ_i	$\phi\mu_i + a_i\mu_i^2$
Poisson	$N\left(0,\sigma_u^2\right)$	$Gamma\,(1/\alpha,\alpha)$	μ_i	$\mu_i + a_i\mu_i^2 + \alpha\exp\left[\sigma_u^2\right]\mu_i^2$

Note: $a_i = \exp\left[\sigma_u^2\right] - 1$

follows a Poisson distribution, with $E\left(z_i|X_i,u_i,v_i,\right) = \lambda_i*$, $\log\left(\lambda_i*\right) = X_i\beta + u_i + \log\left(v_i\right)$, $u_i \sim \left(0,\sigma_u^2\right)$, $v_i \sim Gamma\,(1/\alpha,\alpha)$ and $u_i \perp v_i$. Integrating out the random effects v_i from the joint distribution of $\left(z_i,u_i,v_i\right)|X_i$ we can show that this model is equivalent to a negative-binomial model for $z_i|\left(X_i,u_i\right)$. It gives $E\left(z_i|X_i,u_i\right) = X_i\beta + u_i = \lambda_i$ (say) and $Var\left(z_i|X_i,u_i\right) = \lambda_i + \alpha\lambda_i$ (Lee and Nelder , 2001). The marginal variances implied by these models are listed in Table 1.

We can infer from Table 1 that all three models, i.e., Poisson-normal HGLM, quasi- Poisson-normal HGLM, and Poisson-normal-gamma HGLM, give the same marginal mean but different marginal variances. For $\phi = 1$, the quasi-Poisson-normal HGLM becomes the Poisson-normal HGLM. If $\phi > 1$, we have extra overdispersion, whereas $\phi < 1$ $z_i|u_i$ has an underdispersion, leading to a different correlation structure than does Poisson-normal HGLM. The negative-binomial-normal (or Poisson-normal-gamma) model (with $\sigma_u^2 > 0$ and $\alpha > 0$) simply allows for an excessive overdispersion than does the Poisson-normal model.

3 Estimation and model selection with HGLM

For inferences about HGLMs from the models described in Section 2, where in matrix notation,

$$g\left(\lambda\right) = X\beta + Z\mathbf{w} \tag{9}$$

where $E\left(\mathbf{z}\right) = \lambda$, $\mathbf{z}^T = \left(z_1,z_2,\ldots,z_n\right)$, X is the model matrix for the fixed effects β, $\mathbf{w}^T = \left(w_1,w_2,\ldots,w_n\right)$ is the vector of random effects, and Z is the model matrix associated with $\mathbf{w}$. We use the h-likelihood (hierarchical likelihood) presented by Lee and Nelder (1996) as

$$h = \sum_{i=1}^{n}\log f\left(z_i|X_i,w_i\right) + \log f\left(\mathbf{w}\right) \tag{10}$$

For estimation of the parameters in spatial covariance θ (consisting of Σ, ϕ, and α), we maximize the adjusted profile likelihood

$$p_{\beta,\mathbf{w}}\left(\theta\right) = \left(h - \frac{1}{2}\log|\frac{\partial^2 h}{\partial\left(\mathbf{w},\beta\right)\left(\partial\mathbf{w},\beta\right)^T}|\right)_{\mathbf{w}=\hat{\mathbf{w}},\beta=\hat{\beta}} \tag{11}$$

which gives an extension of restricted maximum (log-)likelihood (Lee et al. , 2006).

Because we consider spatial data where there are as many random effects as the number of observations, a simple penalized quasi-likelihood (Breslow and Clayton , 1993) introduces serious bias in the parameter estimates (see further discussion in Lee and Lee , 2011). Thus, in this paper, we use the extended Fisher's scoring algorithm described in Lee and Lee (2011) , with necessary modification to handle a quasi-Poisson model, for fitting generally structured spatial models. An R (R Development Core Team , 2011) implementation of the above algorithm can be obtained from the authors (we also plan to make it available on CRAN). In the following subsection, we discuss the model selection techniques used with hierarchical generalized linear models.

3.1 Model selection

For model selection, Lee et al. (2006) proposed various h-likelihood-based statistics, e.g., for estimation and model selection of variance and covariance models, the adjusted profile h-likelihood $p_{\beta,\mathbf{w}}()$ defined in (11) is suggested; for the estimation of fixed effects, β, $p_{\mathbf{w}}$ is suggested. For nested models, differences in $-2p_{\mathbf{w}}()$ can be treated the same way as is done for the classical likelihood ratio test. For nonnested models we can use information criteria such as the conditional Akaike's information criteria (cAIC), based upon the first component of h-likelihood, which is given as

$$cAIC = d + 2p_d \tag{12}$$

where $d = -2\left(l\left(\hat{\lambda}; \mathbf{z}|\mathbf{w}\right) - l\left(\mathbf{z}; \mathbf{z}|\mathbf{w}\right)\right)$, $l\left(\hat{\lambda}; \mathbf{z}|\mathbf{w}\right) = \log f\left(\mathbf{z}|\mathbf{w}; \hat{\lambda}\right)$, and $n - p_d$ is the degrees of freedom for the deviance d (Lee et al. , 2006), with $p_d = trace\left(H^{-1}H^{*}\right)$; $H = -2\frac{\partial h^2}{\partial(\beta,\mathbf{w})\partial(\beta,\mathbf{w})^T}$ and

$$H^{*} = \begin{bmatrix} X^TWX & X^TWZ \\ Z^TWX & Z^TWZ \end{bmatrix}$$

where X and Z are as in equation (9) and W is the diagonal GLM weight matrix (McCullagh and Nelder , 1989), whose i^{th} diagonal element is given by $W_{i,i} = \left(\frac{\partial^2\eta}{\partial\lambda_i^2}\right)^2 V^{-1}\left(\lambda_i\right)$ where V is the GLM variance function (McCullagh and Nelder , 1989). Spiegelhalter et al. (2002) refer to the cAIC as the deviance information criterion in connection to selection of Bayesian models, whereas Donohue et al. (2011) use it for frequentist model selection.

4 Analysis of reindeer pellet-group counts

We analyze a real data set on reindeer fecal pellet-group counts. The data were obtained through a survey conducted at Storliden Mountain (504 m MLS; N 65° 13′, E 18° 53′) in northern Sweden. The size of the study area was 25 km^2, and in the center of the area eight windmills were built 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 during May to October and are only gathered for the marking of the calves at the beginning of July. The survey was conducted with a point transect survey 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 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 registered, and the center of each plot was marked with an orange wood stick.

The pellet groups were counted by the technique of fecal standing crop (FSC) in 2009 and fecal accumulation rate (FAR) in 2010. A pellet group was counted against a certain plot if the center of the group was found inside the plot. As an animal might move as it defecates, the pellets could spread over a large area. Therefore, a pellet group was defined by a cluster of 20 or more pellets.

4.1 Spatial modeling of reindeer pellet-group counts

In order to model preference of reindeer grazing area, we model the pellet-group count within each year. 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 feature indicates (possible) inappropriateness of the ordinary count data models, e.g., the Poisson GLM. Keeping this in mind, we apply the three different models for overdispersion seen in Table 1, where the spatial covariance structure for the normal random effects are either $\Sigma = \tau\left(I - \rho D\right)$ (i.e., CAR, which includes Poisson-normal HGLM as a special case for $\rho = 0$) or $\Sigma = \sigma^2\left(I + D\right)\left(I + D\right)^T$. We also fit a Poisson GLM, as it would be done in regression krigging. A detailed list of the fitted models is given in Table 2.

Since the generalized linear models are uniquely specified by the mean (in our case, λ_i) and the variance function, $V\left(\lambda_i\right)$, we use only these parameters and functions to specify the models. All the spatial models presented in Table 2 can be fitted by using HGLM algorithm (an R implementation is made available as supplementary material). For the 2009 data, we started with a large (full) model containing 14 covariates: the (log-)distance from the power grid, (log-)distance from the nearby main road, slope of the location, ruggedness index, elevation, forest age structure, dummies (1/0) for clear-cuts, young forest, coniferous forest, broad-leaved forest, flat area, south-east slope, north-west slope, and north-east slope. For 2010 data, we excluded the dummies for broad-leaved forest from the full model. It was noted that in the 2010 data,

Table 2: Specifications of the fitted models

Model	Description	Mean ($g(\lambda_i) = \eta_i$)	Variance function ($V(\lambda_i)$)	Random effects and their distributions
Model I	Poisson GLM	$\log(\lambda_i) = X_i\beta$	λ_i	No
Model II	Poisson-normal HGLM	$\log(\lambda_i) = X_i\beta + u_i$	λ_i	$\mathbf{u}^T = (u_1, u_2, \ldots, u_n)$, $u_i \sim N(0, \tau)$
Model III	Overdispersed Poisson-normal HGLM	Model II	$\phi\lambda_i$	Model II
Model IV	Poisson-normal HGLM with CAR	Model II	Model II	$\mathbf{u} \sim N(\mathbf{0}, \Sigma)$, $\Sigma = \Sigma = \tau(I - \rho D)$
Model V	Poisson-normal-gamma HGLM with CAR	$\log(\lambda_i) = X_i\beta + u_i + v_i$	Model II	$\log(w_i) = v_i$, $w_i \sim Gamma(\alpha, 1/\alpha)$, $\mathbf{u}$ is as in Model IV
Model VI	Overdispersed Poisson-normal HGLM with CAR (QCAR)	Model II	Model III	Model IV

Table 3: Model fit statistics with full set of covariates

Model	Fit Statistics for 2009 data			Fit Statistics for 2010 data		
	-2ML[1]	-2RL[2]	cAIC[3]	-2ML	-2RL	cAIC
Model I	600.05	-	630.05	405.26	-	433.26
Model II	577.01	620.38	588.12	389.01	436.13	402.65
Model III	557.01	600.34	548.76	265.67	311.43	226.21
Model IV	576.61	617.79	586.75	392.96	434.52	405.84
Model V	579.23	591.86	576.82	389.92	395.88	371.56
Model VI	547.26	585.56	524.64	248.35	291.85	200.57

[1] -2ML is $-2 \times \log(\text{likelihood})$ for Model I, and it is $-2p_u(h)$ for Models II-VI.

[2] -2RL is $-2p_{\beta,u}(h)$ which is not meaningful for Model I hence not reported.

[3] cAIC is the AIC for Model I.

broad-leaved forests had only 0 pellet-group count, which indicates that special care is necessary to tackle the exploding tendency of the ML estimate, if it exists, of the respective parameter (Feinberg and Rinaldo , 2007). However, we found that the MLE of the other parameters does not change a lot after broad-leaved forest is dropped from the model. Therefore, we dropped it from the full model.

Table 3 reveals that the quasi-Poisson-normal HGLM with $\phi < 1$ and CAR (QCAR) specification (Model VI) fits the data best, as it had the highest adjusted log-(profile)likelihood and lowest cAIC. The models with covariance structure defined in Equation (7) could not produce a better fit than the QCAR model. Therefore, we do not report those results in this paper.

With the best-fitted full model in hand, we gradually delete covariates one at a time from the model on the basis of the absolute t-value (lowest t-value deleted first) 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 2009 and 2010 data are presented in Table 4. From the results (Table 4), 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 ¡ 0.001 for both GLM and QCAR). Its positive coefficient estimate reveals that the pellet-group counts were higher at places farther away from power grid lines.

A plot of the observed responses against the fitted values (for 2009 FSC count) are given in Figure 1. The same plots for the 2010 FAR count data reveal the same overall pattern. Therefore, those plots are not shown in this paper however,the plot for the final QCAR model with selected of covariates is presented in Figure 2 along with the fit of a hurdle model, computed via MCMC, in order to provide a comparison between HGLM and hurdle model.

From Figure 1 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 Figure 1, we see that QCAR (lower right in Figure 1; which is also the best fit model in terms of cAIC, see Table 3) not only improves the mean prediction, it also reduces the prediction error variance.

Comparing the four plots of the observed counts against the in-sample predictions (i.e. fitted values; Figure 1) we see that simple Poisson GLM (upper left) gives very poor fits, especially for the extreme values (counts 0 and 5). By including random effects (Poisson-normal HGLM, upper right plot in Figure 1), we can improve the fit of the model. However, the simple Poisson-normal HGLM (lower left plot in Figure 1) was not enough for modeling excessive zero counts. By modeling spatial correlation (Poisson-normal HGLM with CAR), we get a better fit compared to the Poisson-normal HGLM with independent random effects. We also found that the excessive overdispersion by assuming negative binomial-normal HGLM with CAR does not improve the fit. Finally, QCAR gives the best fit. The same findings hold for both 2009 and 2010 data. Here, the estimate of ϕ is less than 1, so that underdispersion of $z_i|u_i$ helps with a better fit for the data.

Table 4: Estimated model parameters and fit statistics of final Poisson GLM and quasi-Poisson-normal HGLM with CAR (QCAR)

Parameters	For 2009 FSC Counts		For 2010 FAR Counts	
	GLM	QCAR	GLM	QCAR
Intercept	-18.916	-18.171	-12.210	-10.141
	(3.782)	(4.491)	(4.562)	(3.877)
Northwest slopes	-0.489	0.656		
	(0.304)	(0.364)		
Southeast slope			0.696	1.088
			(0.316)	(0.381)
Elevation	0.007	0.007	-0.005	-0.005
	(0.002)	(0.003)	(0.003)	(0.003)
Distance to power lines	1.897	1.728	1.569	1.238
	(0.426)	(0.519)	(0.499)	(0.451)
Clear-cuts			0.607	
			(0.346)	
τ		1.324		1.443
		(0.270)		(0.692)
ϕ		0.737		0.332
		(0.082)		(0.035)
ρ		3.038		3.175
		(0.750)		(0.262)
-2ML	610.40	558.084	408.78	23.45
-2RL		575.893		41.16
AIC†	618.40	526.254	418.78	47.90

Notes: Values in parentheses represent std. err. ML (maximum log-likelihood for GLM); it is $-2p_u\,(h)$ for QCAR; $-2RL = -2p_{u,\beta}\,(h)$.

†AIC is conditional AIC for QCAR.

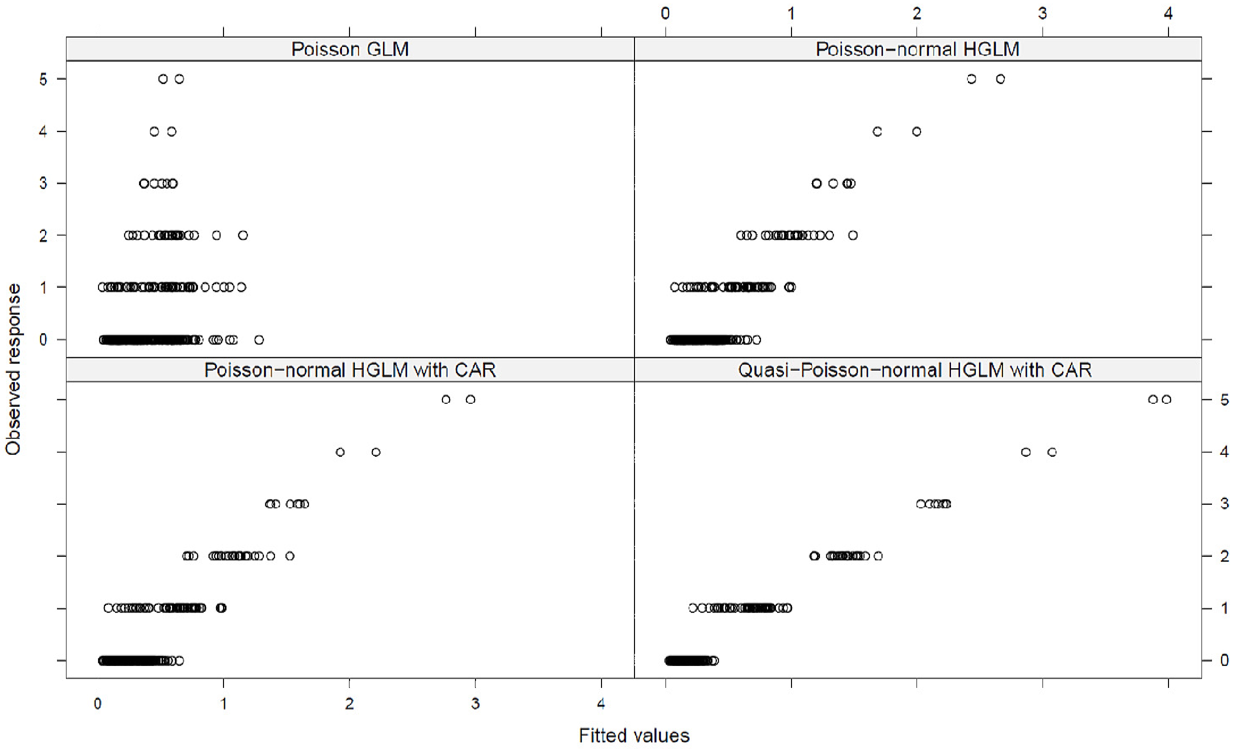

Figure 1: Plots of observed vs. fitted values for different models (with full set of covariates) with 2009 FSC count.

4.2 Comparison with hurdle model

For count response with excess zeros, hurdle model is frequently used. Therefore, we also analyze the reindeer pellet group counts by using a hurdle model. For the hurdle model we use the following model specification (Equation (13)).

$$\left.\begin{array}{l} P\left(y_i>0\right)=p_i ; y_i \sim Bin\left(p_i\right), logit\left(p_i\right)=X_i \beta \\ P\left(y_i=k \mid u_i\right)=p_i \frac{\lambda_i^k \exp \left[-\lambda_i\right]}{k!(1-\exp [-\mu])}, k=1,2 \ldots ; \log \left(\lambda_i\right)=X_i \beta+u_i \\ \mathbf{u} \sim N(\mathbf{0}, \Sigma) \end{array}\right\} \tag{13}$$

where Σ has a CAR specification as in Model IV (see Table 2). Model (13) gives $E\left(y_i \mid u_i\right)=p_i \frac{\lambda_i}{1-\exp \left[-\lambda_i\right]}$.

Computational procedure for Model (13) under non-Baysian approach is not available, up to date. Therefore, we carry out the model computation in Bayesian way by using WinBugs (Spiegelhalter et al. , 2007) running from R (Sturtz et al , 2005) in a similar way as it is in Neelon et al. (2012). Weekly informative priors (e.g. $N(0,1000)$ and $Uniform\,(0,100)$ for the parameters in $(-\infty, \infty)$ and $(0, \infty)$ respectively) were used so that the results are comparable with those in Section 4.1[2].

For the binary model in Equation (13) we use 7 covariates: South-east slope, Young forest, Clear cuts, forest age structure, Elevation and log-distance to power grid. These variables were selected through a separate binomial model for $P\,(Count>0)$ based on lowest AIC. For the truncated Possion model in Equation (13) we use the same set of covariates as in Table 4.

The median of psoterior distribution computed from 5001 MCMC samples (10000 MCMC simulations and 50% burned-in over 3 chains; every 3rd sample is retained from each chain) on the parameters are used to compute fitted values and their median is reported for the Bayesian models in Figure 2 (hurdle, Binary and trnc. Poisson). Top panel of Figure 2 show the fit of QCAR (left) and hurdle (right). The bottom panel shows the fit of the Binary (right) and the truncated Poisson (left) part in the hurdle model. A small jitter was used for the figures in right column to make numerically equivalent values visible in the plot. Figure 2 shows that the HGLM model with QCAR (Model VI) provided better fit than Bayesian hurdle

[2] Bugs program file available at: *http://users.du.se/∼maa/EES2012/*); alternatively e-mail to *maa@du.se*

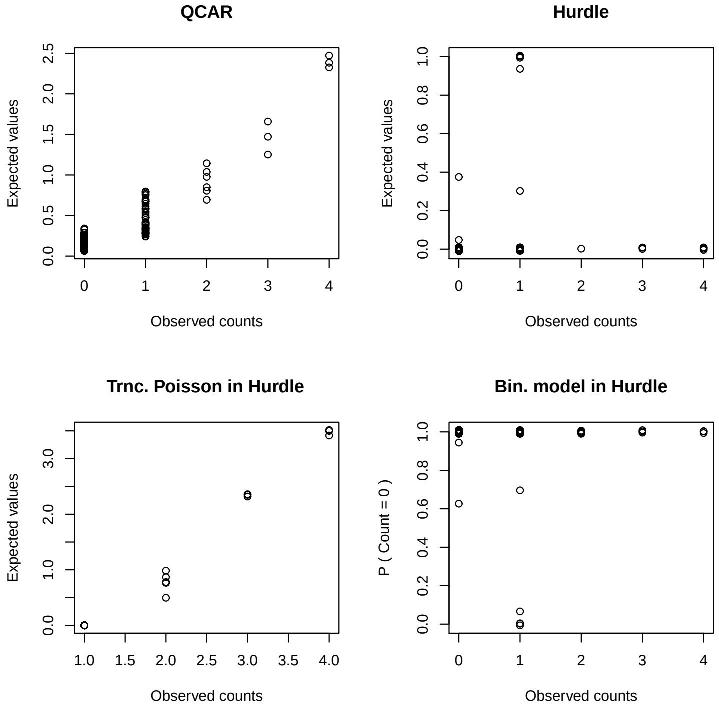

Figure 2: Plots of observed vs. fitted values for QCAR, hurdle and the individual components of the hurdle model for 2010 FAR counts.

model. Though the truncated Poisson part in the hurdle model did a great job, the failure of the Binary part downgraded the overall prediction. We tried will all the available variables in hand to improve the performance of the binary model we failed. We could not fit a spatial correlation in the binary part either because WinBugs failed to fit such model. The above failure can be due to the lack of information on covariance parameters as we have only one binary response for each plot.

5 Concluding discussion

In this paper, we demonstrated a method for the joint modeling of the mean and covariance structure for spatially correlated count response, including excessive zero counts, by using HGLM. We provided intuitive explanations for the proposed spatial covariance structures, outlined the estimation of the model parameters, and discussed the techniques of model selection via the h-likelihood method. The HGLM framework provides us with unified tools for model fitting, model comparison, and drawing model parameter inferences.

By analyzing a real data set on reindeer pellet-group counts, we showed that a Poisson GLM by ignoring spatial correlation can lead to poor model fit. Such a simplified model produces bad residuals (due to the lack of fit), which lead to wrong conclusions about the spatial correlation. Consequently, the regression kriging prediction based upon those residuals, which is often suggested in the literature (see, e.g. Cressie , 1993; Bivand et al. , 2008), might produce poor spatial prediction.

From the results of the fitted quasi-Poisson-normal HGLM (see Table 4), we conclude that several environmental variables, e.g., slope, elevation, and vegetation type of the location as well as human development activities, e.g., power lines, are significant factors for explaining reindeer habitat preference.

In the literature, hurdle models are widely used for analyzing count response with excessive zeros. However, hurdle models do not allow an underdispersion with excessive zeros. In practice such data sets often exist, for example an incidence rate of hospitalization and Tin (2008) pointed out there is no model available to fit these data. The quasi-Poisson-normal HGLM allow underdispersion ($\phi < 1$) with excessive zeros for both 2009 and 2010 data. The spatial hurdle model allows only an extra overdispersion when zero-inflation occurs than does Poison-normal HGLM. Therefore, by analyzing a real data set, we show that a hurdle model may not necessarily be a better choice than a HGLM model with similar linear covariates and correlation structures. It is interesting future work to extend hurdle models to allow underdispersion with excessive zeros. Moreover, computation of the spatial hurdle models are done by using Bayesian MCMC techniques which are computationally too intensive. The HGLM model computed via h-likelihood method (an R is made available) provides much fasted model fitting than the MCMC methods.

References

Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environ Ecol Stat, 9(4):341–355.

Bivand RS, Pebesma EJ, Gomez-Rubio, V (2008) Applied Spatial Data Analysis with R. Springer, New York.

Breslow NE, Clayton, DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc, 88:9–25.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to Distance Sampling–Estimating Abundance of Biological Populations. Oxford University Press, New York.

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

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

Cressie NAC (1993) Statistics for Spatial Data (revised). Wiley, New York.

Dolan DM, El-Shaarawi AH, Reynoldson TB (2000) Predicting benthic counts in Lake Huron using spatial statistics and quasi-likelihood. Environmetrics, 11:287–304.

Donohue MC, Overholser R, Xu R, Vaida F (2011) Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika, doi: 10.1093/biomet/asr023.

Edge WD, Marcum CL (1989) Determining elk distribution with pellet-group and telemetry techniques. J Wildl Manage, 53:621–624.

Efron B (1986) Double exponential families and their use in generalized linear models. J Am Stat Assoc, 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. Environ Ecol Stat, 18:291–314.

Feinberg SE, Rinaldo A (2007) Three centuries of categorical data analysis: log-linear models and categorical data analysis. J Stat Plan Inference, 137:3430–3445.

Gribko LS, Hohn ME, Ford WF (1999) White-tailed deer impact on forest regeneration: modeling landscape-level deer activity patterns. In: Stringer JW, Loftis DL (eds.) Proceedings, 12th Central Hardwood Forest Conference. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, pp 178–185.

Kindermann R, Snell JL (1980) Markov random fields and their applications. Contemporary Mathematics 1, American Mathematical Society, Providence, RI.

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

Lee W, Lee Y (2011) Modifications of REML algorithm for HGLMs. Stat Comput, doi: 10.1007/s11222-011-9265-9.

Lee Y, Nelder JA, Pawitan Y (2006) Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Chapman & Hall/CRC, Boca Raton.

Lee Y, Nelder JA (2001) Two ways of modeling overdispersion in non-normal data. Appl Stat, 49:591–598.

Lee Y, Nelder JA (1996) Hierarchical generalized linear models (with discussion). J R Stat Soc Series B Stat Methodol, 58:619–656.

Lichstein JW, Simon TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in Ecology. Ecol Monogr, 72:445–462.

McCullagh P, Nelder JA (1989) Generalized Linear Models. Chapman & Hall, London.

Molenberghs G, Verbeke G, Demetrio CGB (2007) An extended random-effects approach to model repeated, overdispersed count data. Lifetime Data Anal, 13:513–531.

Neelon B, Ghosh P, Loebs FP (2012) A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. J R Stat Soc Ser A, 175:1–25.

Neff DJ (1968) The pellet-group count technique for big game trend, census, and distribution: a review. J Wildl Manage, 32:597–614.

Nelder JA, Pregibon D (1987) An extended quasi-likelihood function. Biometrika, 74:221–232.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Skarin A (2007) Habitat use by semi-domesticated reindeer, estimated with pellet-group counts. Rangifer, 27:121–132.

Spiegelhalter DJ, Thomas A, Best N, Lunn D (2003) Winbugs Version 1.4: Users Manual. MRC Biostatistics Unit, Cambridge, URL http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf (last accessed: Aug. 25, 2012).

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measure of model complexity and fit (with discussion). J R Stat Soc Series B Stat Methodol, 64:583–640.

Sturtz S, Ligges U, Gelman A (2005) R2WinBUGS: a package for running WinBUGS from R. J Statist Software, 12:1–16.

Tin A (2008) Modeling Zero-Inflated Count Data with Underdispersion and Overdispersion. SAS Global Forum 2008 Statistics and Data Analysis.

Xiaoping J, Banarjee S, Carlin, BP (2007) Order-free co-regionalized areal data models with application to multiple-disease mapping. J R Stat Soc Series B Stat Methodol, 69:817–838.