HÖGSKOLAN
DALARNA

MASTER THESIS IN MICRODATA ANALYSIS

Computation and application of the spatial zero inflated count response

Author:
Shengqiang Guo

Supervisor:
Md. Moudud Alam

June, 2014

Business Intelligence Program
School for Technology and Business Studies
Dalarna University

Abstract

Generalized linear models and its extensions are widely used for analyzing non-normal data. But Poisson mixed model may exhibit inadequate fitting and inference when encounter excessive zero counts. Mixed hurdle model is a preferable method to solve the problem. Nevertheless, it is still a challenge to use the mixed hurdle model to deal with correlated data. There are a few computational procedure for hurdle model can be used to calculate, particularly for the model with random effects being correlated between non-zero and zero response parts. In our paper we display a method to fit the hurdle model with conditionally autoregressive random effects for the spatial data. Based on the extended algorithm, some modifications are made to the existing procedure in R to help us to fit the data. We conduct Monte-Carlo simulation to study the finite sample properties of our model. The result shows that the new procedure fit the model well. The estimation becomes better with the increase of measurement in each subject. At last, we apply the new procedure to a real problem. The dataset is about reindeer spatial distribution related to the wind power establishments at Storliden Mountain in North Sweden. The new procedure gives a better fit of the real problem than a usual Poisson mixed model.

Keywords: Zero-inflated data, Hurdle model, CAR random effects, Reindeer pallet group distribution.

CONTENTS

1. Introduction

Generalized linear models (GLM) and its extensions are widely used to deal with the data which do not follow the normality assumption. But in real applications, the situation with surplus zeros in response often happens, especially for count dependent variable. Zero-inflated Poisson (ZIP) regression and Hurdle model are two available methods for such problem (Lambert, 1992). They have been applied in health medical care (Winkelmann, 2004), economics, banking (Moffatt, 2005) and many other areas. In a hurdle model, the data was separated into non-zero and zero parts. They are analyzed with different approaches.

Min and Agresti (2005) exploited nonparametric method and normal distribution to the random effects. The gauss-hermite quadrature technique for calculating marginal likelihood was used to estimate the parameters. But this approach computationally slowly. What's more, the accuracy decrease with the increasing of the random components' number. Molas and Lesaffre (2010) has extended the h-likelihood estimation to the hurdle model with random effects to handle the uncorrelated data. They also took advantage of two datasets to examine the method.

Zero-inflated binomial (ZIB) model was also introduced based on the ZIP method. The within-subject correlation and the correlation between different subjects has been considered in the random effects to cope with the repeat measurements (Hall, 2000). Although there are many literatures involve in the hurdle model, few of them considered its application in spatial data. For a spatial data analysis, correlations always exist among different observations. What's more, there are few computational tools to handle with the hurdle model in spatial data manipulation which makes it is inconvenience for daily analysis.

In this paper, we try to extend the hurdle model with random effects which are correlated within the zero and non-zero parts. A new procedure was implemented in R software to handle the mixed hurdle model. In order to deal with the spatial data, we assume that the random effects follow a Gaussian conditionally autoregressive process. We illustrate that the hurdle mixed model, with correlated and independent random effects, can be fitted with iteratively reweighted least squares (IWLS) algorithm. A simulation study is conducted to understand the finite sample properties of the procedure. Then we apply the procedure to a real data problem of modeling the reindeer distribution on the Storliden Mountain in north Sweden.

Section 2 describes the statistical models and the proposed estimation technique. The numerical aspects of mixed hurdle model was also represent in this review. The simulation study is presented in Section 3. Real example analysis is shown in Section 4. Different types of models are considered in comparison. A discussion is presented in Section 5.All the R codes in use are provided in the Appendix.

2. Methodology

Three kinds of models GLM, GLMM and hurdle model are used to fit the reindeer dataset. For the hurdle model, we derive a new estimation technique based on hierarchical likelihood approach. We consider both the mixed model without correlated and with correlated random effects inside the zero and non-zero part.

2.1 Generalized linear model

There are many limitations when we apply the linear model in the real data analysis, such as the linearity, normality and homoscedasticity. The GLM provides an approach which relax these assumptions.

Assuming $\boldsymbol{y}$ follows the distribution which belongs to the exponential family (including binomial, Poisson, normal etc.). For the GLM, we have:

$$\begin{aligned} &\boldsymbol{\mu}=\mathrm{E}(\boldsymbol{y}) \\ &g(\boldsymbol{\mu})=\boldsymbol{\eta}=\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{1}$$

Where, $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{y}$, $\boldsymbol{X}$ is the design matrix of fixed effects, $\boldsymbol{\beta}$ is the regression coefficient, $g(.)$ is the link function which must be differentiable and monotone. Take the exponential family in a general way for a single observation:

$$f(y;\theta,\phi)=\exp\left\{\frac{(y\theta-b(\theta))}{a(\phi)}+c(y,\phi)\right\} \tag{2}$$

θ is a function of the distribution's location parameter, the $a(.)$, $b(.)$, $c(.)$ are some certain functions. Then maximum likelihood (ML) method can be used to estimate the parameter of the GLM. The log likelihood for only one observation can be represented as:

$$l=\log p(y)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi) \tag{3}$$

Where $\boldsymbol{\beta}$, the regression coefficients of the model is related to the parameter $\boldsymbol{\theta}$**.** Take $\frac{\partial\mu}{\partial\theta}=V$, $\left(\frac{d\eta}{d\mu}\right)^2\cdot V=W^{-1}$, according to the chain rule to the jth regression coefficient β_j we have:

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \frac{\partial l}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j} \\ &= \frac{(y-\mu)}{a(\phi)}\frac{1}{V}\frac{d\mu}{d\eta}x_j \\ &= \frac{W}{a(\phi)}(y-\mu)\frac{d\eta}{d\mu}x_j\end{aligned} \tag{4}$$

Then the likelihood for only one regression parameter can be illustrated as:

$$\sum_i \frac{W_i(y_i-\mu_i)}{a(\phi)}\frac{d\eta_i}{d\mu_i}x_{ij} = 0 \quad , i = 1,2,...,n \tag{5}$$

This simplification enlightens us to use the iteratively reweighted least squares (IWLS) algorithm to obtain the maximum likelihood estimate (Olsson, 2002).

2.2 Generalized linear mixed model

If we have limited number of observations for each parameter, but large number of parameters to estimate, the GLM model might have some limitations to deal with the problem (Pawitan, 2001). For instance, we have 50 subjects to explain with only two observations for each of them. In addition, there is a general assuming in GLM as we did in linear model is all the response observations are independent. However, these observations might be correlated in reality (Jiang, 2007). The necessity to consider GLMM increases when we confront these matters. Random effects are added in GLMM to describe the problems. To demonstrate,

$$\begin{aligned}&\boldsymbol{\mu}\text{=E}(\boldsymbol{y}\,|\,\boldsymbol{b}) \\ &g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zb}\end{aligned} \tag{6}$$

Where, $\boldsymbol{\mu}$ is the conditional expectation of $\boldsymbol{y}$ given $\boldsymbol{b}$, $\boldsymbol{\beta}$ is the regression coefficient for fixed effects, $\boldsymbol{b}$ illustrates the random effects, often with $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{G})$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the known design matrix for fixed effects and random effects, $g(.)$ is the link function as in GLM. The joint likelihood of the GLMM is:

$$l = \log p(\boldsymbol{y}\,|\,\boldsymbol{b}) + \log p(\boldsymbol{b}) \tag{7}$$

The first term on the right hand is the same as we describe before, $\boldsymbol{b}$ follows the normal distribution which also belongs to the exponential family. To estimate the parameters of the model, some approximation are needed. (Pawitan, 2001; Jiang, 2007; McCulloch & Searle, 2001)

2.3 Hurdle model

We often fit the counts response with Poisson distribution in application. But the model may fit poorly encountering excessive zeros in the response. Hurdle model is a suitable technique to account for the problem. The approach applies a binomial distribution to manage whether the outcome is zero or not, and then a truncated Poisson model is used for the non-zero part (McCulloch & Searle, 2001). We usually take a logit link to describe the binary model and a log link for the truncated Poisson part.

In a hurdle model, for the *i*th subject (*i*=1,2,…,*N*), *j*th measurement (*j*=1,2,…,n_i) the probability can be showed as:

$$\begin{aligned} P(Y_{ij}=0) &= 1-p_{ij} \quad , \\ P(Y_{ij}=k) &= p_{ij}\cdot\frac{\mu_{ij}^{k}}{k!\cdot(\exp(\mu_{ij})-1)} \end{aligned} \tag{8}$$

The p_{ij} is the probability that it is not zero for one certain observation. μ_{ij} is the mean of the non-zero part in not truncated Poisson distribution. The complete Poisson hurdle model with random effects given in matrix manner:

$$\begin{aligned} logit(\boldsymbol{p}) &= \boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{Z}_1\boldsymbol{b}_1 \\ \log(\boldsymbol{\mu}) &= \boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{Z}_2\boldsymbol{b}_2 \end{aligned} \tag{9}$$

If the random effects $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are independent with each other, then the two sections above can be handled separately. We can treat the Bernoulli part in a same way as in GLMM. However, hurdle model requires some adjustment to the GLMM procedure which we describe in the following subsection.

Let the conditional response on random effects follows a modified exponential distribution:

$$f(y_{ij}\mid b_i;\theta_{ij},\phi)=\exp\left\{\frac{(y_{ij}\theta_{ij}-b(\theta_{ij})-\log[T(\theta_{ij})])}{a(\phi)}+c(y_{ij},\phi)\right\} \tag{10}$$

Where $a(\phi)=1$, $\theta_{ij}=\log(\mu_{ij})$, $\mathrm{b}(\theta_{ij})=\exp(\theta_{ij})$, $T(\theta_{ij})=1-\exp(-\exp(\theta_{ij}))$.

If the random effects are independent within the truncated Poisson part, Molas and Lesaffre (2010) showed a modified algorithm to estimate the parameters in the model. But they may be correlated in reality, especially in spatial data analysis. Assuming the random effects in truncated Poisson part follows:

$$\boldsymbol{b}\sim N(\boldsymbol{0},\Sigma) \tag{11}$$

Where $\sum = \tau(\boldsymbol{I} - \rho \boldsymbol{D})^{-1}$, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{D}$={d_{st}} is a neighborhood matrix in spatial data analysis. The elements d_{st} equals to 1 when the locations s and t are neighbors and all the diagonal elements are 0 in $\boldsymbol{D}$.

Let ω_i to be the ith eigenvalue of the neighborhood matrix $\boldsymbol{D}$, $\boldsymbol{V}$ is the corresponding matrix whose column vectors are orthonormal, λ_i is the ith eigenvalue of the matrix $\boldsymbol{\Lambda}$. We have,

$$\Sigma^{-1} = \frac{(\boldsymbol{I} - \rho \boldsymbol{D})}{\tau} = \boldsymbol{V \Lambda V}^T = \boldsymbol{V} diag\{\lambda_i\} \boldsymbol{V}^T = \boldsymbol{V} diag\{\frac{1-\rho\omega_i}{\tau}\} \boldsymbol{V}^T \tag{12}$$

Take $\tilde{\boldsymbol{Z}} = \boldsymbol{ZV}$, $\boldsymbol{b}^* = \boldsymbol{V}^T \boldsymbol{b}$, then

$$\begin{aligned} &\boldsymbol{b}^* \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}) \\ &\log(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{b}^* \end{aligned} \tag{13}$$

For different i, j, we have $b_i^* \perp b_j^*$ and $V(b_i^*) = 1/\lambda_i$. Now we can treat the model as uncorrelated situation above and fit the model in R. (Alam, et al., 2014).

2.4 Estimation of hurdle model

In the above section, we showed the truncated Poisson model follows a modified exponential form;

$$f(y_{ij} \mid b_i; \theta_{ij}, \phi) = \exp\left\{ \frac{(y_{ij}\theta_{ij} - b(\theta_{ij}) - \log[T(\theta_{ij})])}{a(\phi)} + c(y_{ij}, \phi) \right\} \tag{14}$$

Where $a(\phi) = 1$, $\theta_{ij} = \log(\mu_{ij})$, $\mathrm{b}(\theta_{ij}) = \exp(\theta_{ij})$, $T(\theta_{ij}) = 1 - \exp(-\exp(\theta_{ij}))$,

and $\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Zb}$. Then the log likelihood for the truncated Poisson:

$$\begin{aligned} l &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log(f(y_{ij} \mid b_i; \theta_{ij}, \phi)) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{(y_{ij}\theta_{ij} - b(\theta_{ij}) - \log[T(\theta_{ij})])}{a(\phi)} + c(y_{ij}, \phi) \end{aligned} \tag{15}$$

The score equation of l with respecting $\boldsymbol{\beta}$ in matrix form:

$$\begin{aligned}\nabla &= \frac{\partial l}{\partial \boldsymbol{\beta}_j} = \frac{\partial l}{\partial \boldsymbol{\theta}}\frac{d\boldsymbol{\theta}}{d\boldsymbol{\mu}}\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}_j} \\ &= \frac{\boldsymbol{y}-\boldsymbol{\mu}-\frac{T^{'}(\boldsymbol{\theta})}{T(\boldsymbol{\theta})}}{a(\boldsymbol{\phi})}\frac{d\boldsymbol{\theta}}{d\boldsymbol{\mu}}\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\boldsymbol{X} \\ &= \boldsymbol{X}^{'}(\boldsymbol{y}-\boldsymbol{\mu}-\frac{\boldsymbol{\mu}}{\exp(\boldsymbol{\mu})-1})\end{aligned} \tag{16}$$

The Hessian matrix:

$$\boldsymbol{H} = \frac{\partial^2 l}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} = -\boldsymbol{X}^{'}\text{diag}\{\boldsymbol{\mu}+\frac{\boldsymbol{\mu}(\exp(\boldsymbol{\mu})-\boldsymbol{\mu}(\exp(\boldsymbol{\mu})-1))}{(\exp(\boldsymbol{\mu})-1)^2}\}\boldsymbol{X} \tag{17}$$

Assuming $\boldsymbol{S} = (\boldsymbol{y}-\boldsymbol{\mu}-\frac{\boldsymbol{\mu}}{\exp(\boldsymbol{\mu})-1})$, $\boldsymbol{W} = \text{diag}\{\boldsymbol{\mu}+\frac{\boldsymbol{\mu}(\exp(\boldsymbol{\mu})-\boldsymbol{\mu}(\exp(\boldsymbol{\mu})-1))}{(\exp(\boldsymbol{\mu})-1)^2}\}$

Then $\nabla = \boldsymbol{X}^{'}\boldsymbol{S}$, $\boldsymbol{H} = -\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X}$.

We can utilize the iteratively reweighted least squares approach to solve the problem. If we define $\boldsymbol{\beta}^{(0)}$ as the initial estimates of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(1)}$ as the estimate of first iteration, according to the Newton-Raphson method (Olsson, 2002), the numerical procedure to estimate $\boldsymbol{\beta}$:

$$\begin{aligned}\boldsymbol{\beta}^{(1)} &= \boldsymbol{\beta}^{(0)} - \boldsymbol{H}^{-1}\nabla = \boldsymbol{\beta}^{(0)} + (\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{S} \\ &\Rightarrow (\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta}^{(1)} = (\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta}^{(0)} + \boldsymbol{X}^{'}\boldsymbol{S} \\ &\Rightarrow \boldsymbol{\beta}^{(1)} = (\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{W}(\boldsymbol{\eta}^{(0)} + \boldsymbol{W}^{-1}\boldsymbol{S})\end{aligned} \tag{18}$$

We will get a series of estimate of $\boldsymbol{\beta}$ with the approximation until it converge. For the hurdle model with random effects, by following Lee and Nelder (1996)

$$\boldsymbol{\varphi} = \begin{pmatrix}\boldsymbol{\beta} \\ \boldsymbol{b}\end{pmatrix},\quad \boldsymbol{X}_h = \begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{O} & \boldsymbol{I}\end{pmatrix},\quad \boldsymbol{\xi} = \begin{pmatrix}\boldsymbol{\eta}^{(0)} + \boldsymbol{W}^{-1}\boldsymbol{S} \\ \boldsymbol{O}\end{pmatrix},\quad \boldsymbol{W}_h = diag(\boldsymbol{W}, Var(\boldsymbol{b})) \tag{19}$$

Where ***X*** is the design matrix for fixed effects, ***Z*** is the design matrix for the random effects, ***O*** is the zero matrix, ***I*** is the identity matrix, *Var(**b**)* is the variance of random effects. The numeric procedure can be present as (see Lee and Nelder, 1996):

$$\boldsymbol{\varphi}^{(1)} = (\boldsymbol{X}_h^{'}\boldsymbol{W}_h\boldsymbol{X}_h)^{-1}\boldsymbol{X}_h^{'}\boldsymbol{W}_h\boldsymbol{\xi}^{(0)} = [(\boldsymbol{W}_h^{1/2}\boldsymbol{X}_h)^{'}(\boldsymbol{W}_h^{1/2}\boldsymbol{X}_h)]^{-1}(\boldsymbol{W}_h^{1/2}\boldsymbol{X}_h)^{'}(\boldsymbol{W}_h^{1/2}\boldsymbol{\xi}^{(0)}) \tag{20}$$

If we treat $\boldsymbol{W}_h^{1/2}\boldsymbol{X}_h$ as the independent variables, and $\boldsymbol{W}_h^{1/2}\boldsymbol{\xi}^{(0)}$ as the response, the estimate of $\boldsymbol{\varphi}^{(1)}$ equals to the result of a simple linear regression. We get a series of estimation of $\boldsymbol{\varphi}^{(n)}$ with iterations until the process has converged.

Take $v_k = \dfrac{\hat{b}_k^2}{1-h_k}$, $\hat{b}_k$ is the estimation of *k*th random effect in the above procedure, *h_k* is the corresponding hat value. Assuming $\theta_0 = \dfrac{1}{\tau}, \theta_1 = -\dfrac{\rho}{\tau}$, as the above section indicate, if we treat ***v***={*v_k*} as the response, and the eigenvalue of neighborhood matrix as linear predictor, a gamma model with inverse link can be used for modelling (Alam, et al., 2014). We can also fit the model with a log link. In section 2.3, we know $V(v_k) = \tau/(1-\rho\omega_k)$. With a log link, the equation can be represent as:

$$\log(Var(v_k)) = \log(\tau) - \log(1-\rho\omega_k) \tag{21}$$

By making a Taylor approximation we have,

$$\begin{aligned}\log(Var(v_k)) &= \log(\tau) - \log(1-\rho\omega_k) \\ &= \log(\tau) - [(-\rho\omega_k) - \frac{(-\rho\omega_k)^2}{2} + \frac{(-\rho\omega_k)^3}{3} + \cdots] \\ &= \pi_0 + s(\omega,\rho)\end{aligned} \tag{22}$$

Where *s*(*ω,ρ*) is a non-linear function of *ω* and *ρ*. A gamma model with log link can be used for modelling where the *s*(*ω,ρ*) can be approximated in a non-parametric way such as by using a cubic spline. Now the parameter can be estimated with a numerical procedure. The simulation study in next section are implemented with inverse link and log link separately. Only τ is considered for the hurdle model with independent random effects. We can take advantage of a gamma model with log link to estimate it.

3. Simulation study

The package *hglm* in R could handle with the Bernoulli part of mixed hurdle model. In this section, we only focus on the truncated Poisson part of hurdle model with random effects. According to the section 2.4, the algorithm is implemented in R software to estimate. We conduct two situations, one with the random effects being correlated and the other with independent random effects. We summarize the result of some Monte-Carlo simulations to assess the sample properties of the estimation. All the programme in use are attached in the appendices.

We conduct the simulation with two group of parameters, each contains an intercept α and two fixed effects parameters β_1, β_2. The number of subjects were set into three types: 50, 100, 200. For each of them, we consider k=2, 5, 10 repeated measurement. All the groups are simulated for 500 times.

If we denote $\hat{\theta}_i$ as the *i*th estimation of parameter θ, the mean estimate of $\hat{\theta}$ can be represent as $\frac{1}{500}\sum_{i=1}^{500}\hat{\theta}_i$ and mean square error (MSE) equals to $\frac{1}{500}\sum_{i=1}^{500}(\hat{\theta}_i-\theta)^2$. We display the estimation and MSE for each parameter in the table. For the correlated random effects simulation, we generated a neighborhood matrix with about 20% of subjects are neighbors. Table 1 display the result of the simulation with inverse link. We represent the outcome of correlated random effect by log link in table 2. The result of independent random effects simulation is presented in table 3. The numbers in parenthesis are the corresponding MSE for each estimation.

Table 1 Simulation result: Truncated Poisson part of mixed hurdle model with correlated random effects (inverse link)

n	k	α=-1.7	β1=1.3	β2=0.45	α=0. 45	β1=0.03	β2=-0.22
50	2	-1.4452	1.2578	0.4524	0.7974	0.0311	-0.2166
		(2.53e-01)	(3.13e-01)	(7.03e-02)	(1.19e+00)	(9.18e-02)	(2.41e-02)
	5	-1.5156	1.3156	0.4436	0.6095	0.0278	-0.2241
		(9.73e-02)	(1.04e-01)	(2.46e-02)	(5.83e-02)	(2.33e-02)	(5.48e-03)
	10	-1.5823	1.3077	0.4533	0.5612	0.0270	-0.2175
		(6.01e-02)	(4.41e-02)	(1.14e-02)	(5.70e-02)	(8.30e-03)	(2.27e-03)
100	2	-1.3967	1.2766	0.4319	0.7466	0.0331	-0.2178
		(1.53e-01)	(1.39e-01)	(3.88e-02)	(5.59e-02)	(4.56e-02)	(9.89e-03)
	5	-1.5223	1.3040	0.4498	0.6784	0.0276	-0.2189
		(5.58e-02)	(4.07e-02)	(1.14e-02)	(1.14e+00)	(9.86e-03)	(2.50e-03)
	10	-1.5888	1.2999	0.4534	0.5837	0.0321	-0.2168
		(3.92e-02)	(2.36e-02)	(5.49e-03)	(5.30e-01)	(4.08e-03)	(1.11e-03)
200	2	-1.3987	1.2920	0.4256	0.7976	0.0266	-0.2160
		(7.74e-02)	(7.05e-02)	(1.77e-02)	(1.35e+00)	(1.81e-02)	(4.70e-03)
	5	-1.5048	1.2923	0.4485	0.6290	0.0265	-0.2194
		(3.99e-02)	(2.28e-02)	(6.46e-03)	(2.00e-02)	(4.60e-03)	(1.19e-03)
	10	-1.5594	1.3022	0.4486	0.5711	0.0335	-0.2196
		(2.77e-02)	(1.13e-02)	(2.49e-03)	(1.14e-01)	(1.95e-03)	(5.42e-04)

Remark: (1) line 1 shows the true parameters; n: No. of subjects; k: No. of measurement in each subject
(2) numbers without parenthesis are the mean estimate; numbers in parenthesis are the corresponding MSE

Table 2 Simulation result: Truncated Poisson part of mixed hurdle model with correlated random effects (log link)

n	k	α=1.7	β_1=1.3	β_2=-0.5	α=0.95	β_1=-0.35	β_2=0.64
50	2	1.7551	1.2907	-0.4972	0.9998	-0.3465	0.6372
		(4.48e-02)	(2.70e-02)	(7.16e-03)	(4.97e-02)	(4.46e-02)	(1.33e-02)
	5	1.7104	1.2996	-0.4952	0.9787	-0.3475	0.6372
		(3.44e-02)	(6.65e-03)	(1.54e-03)	(2.99e-02)	(1.04e-02)	(2.67e-03)
	10	1.7103	1.2970	-0.5012	0.9696	-0.3436	0.6389
		(3.44e-02)	(2.87e-03)	(7.75e-04)	(2.55e-02)	(5.03e-03)	(1.48e-03)
100	2	1.7531	1.2886	-0.4920	1.0063	-0.3384	0.6350
		(3.06e-02)	(1.35e-02)	(3.53e-03)	(2.90e-02)	(1.98e-02)	(5.26e-03)
	5	1.7259	1.3047	-0.5011	0.9848	-0.3507	0.6373
		(2.22e-02)	(2.79e-03)	(8.72e-04)	(2.19e-02)	(5.33e-03)	(1.45e-03)
	10	1.7140	1.2966	-0.4992	0.9779	-0.3518	0.6400
		(2.17e-02)	(1.33e-03)	(3.58e-04)	(1.64e-02)	(2.56e-03)	(6.87e-04)
200	2	1.7415	1.2928	-0.4940	1.0181	-0.3479	0.6332
		(3.16e-02)	(7.08e-03)	(1.54e-03)	(2.76e-02)	(9.99e-03)	(2.95e-03)
	5	1.7200	1.2980	-0.5006	0.9788	-0.3451	0.6391
		(2.29e-02)	(1.57e-03)	(3.47e-04)	(1.99e-02)	(3.17e-03)	(7.35e-04)
	10	1.7157	1.3007	-0.4999	0.9836	-0.3504	0.6395
		(2.67e-02)	(6.68e-04)	(1.82e-04)	(1.78e-02)	(1.14e-03)	(3.49e-04)

Remark: (1) line 1 shows the true parameters; n: No. of subjects; k: No. of measurement in each subject
(2) numbers without parenthesis are the mean estimate; numbers in parenthesis are the corresponding MSE

Table 3 Simulation result: Truncated Poisson part of mixed hurdle model with independent random effects

n	k	α=1.7	β_1=1.3	β_2=-0.5	α=0.95	β_1=-0.35	β_2=0.64
50	2	1.7572	1.2916	-0.5000	1.0239	-0.3442	0.6352
		(3.12e-02)	(2.74e-02)	(6.39e-03)	(4.09e-02)	(4.34e-02)	(9.42e-03)
	5	1.7220	1.2977	-0.4991	0.9822	-0.3440	0.6389
		(2.02e-02)	(6.75e-03)	(1.41e-03)	(1.91e-02)	(1.26e-02)	(3.01e-03)
	10	1.7134	1.3036	-0.4999	0.9681	-0.3530	0.6396
		(1.87e-02)	(3.22e-03)	(6.92e-04)	(1.36e-02)	(4.70e-03)	(1.42e-03)
100	2	1.7470	1.3031	-0.4946	1.0299	-0.3528	0.6355
		(1.70e-02)	(1.36e-02)	(3.61e-03)	(2.40e-02)	(2.05e-02)	(5.34e-03)
	5	1.7259	1.3001	-0.5002	0.9899	-0.3485	0.6384
		(1.09e-02)	(3.38e-03)	(8.00e-04)	(1.03e-02)	(5.88e-03)	(1.53e-03)
	10	1.7122	1.3015	-0.5006	0.9678	-0.3527	0.6398
		(9.10e-03)	(1.45e-03)	(3.80e-04)	(7.25e-03)	(2.53e-03)	(6.62e-04)
200	2	1.7498	1.2953	-0.4935	1.0263	-0.3446	0.6337
		(9.36e-03)	(6.59e-03)	(1.51e-03)	(1.44e-02)	(1.13e-02)	(2.59e-03)
	5	1.7209	1.3020	-0.4987	0.9880	-0.3496	0.6362
		(4.84e-03)	(1.60e-03)	(4.28e-04)	(6.05e-03)	(2.98e-03)	(6.62e-04)
	10	1.7137	1.3011	-0.5000	0.9663	-0.3505	0.6377
		(4.17e-03)	(7.08e-04)	(1.80e-04)	(3.60e-03)	(1.14e-03)	(3.67e-04)

Remark: (1) line 1 shows the true parameters; n: No. of subjects; k: No. of measurement in each subject
(2) numbers without parenthesis are the mean estimate; numbers in parenthesis are the corresponding MSE

According to the simulation result, we can see that with the increasing of measurement in each subject, the estimation becomes better. Table 1 shows the procedure with inverse link capture the variable estimation well. But it is biased for the estimation of intercept. Table 2 indicates the model with log link fits better. This might because that with an inverse link, the result will spread all over the field of real numbers, but the response is only positive. They are not match with each other might lead biased result in numerical procedure, while the log link limits the estimate in a positive domain. Therefore, we use log link to analysis the real problem in next part.

4. Data analysis

We present a data summary before analysis. We start the analysis with a GLM model as Skarin and Rönnegård (2011) did in their paper. The stepwise model selection method was used to achieve the preferable model. Then we maintain the same independent variables while fitting different kinds of models to make a comparison.

4.1 Data summary

The dataset in application has been analyzed in Skarin & Rönnegård (2011). The objective of the data is to test whether the wind power plants have effects on the spatial distribution of reindeer. They only utilized the GLM model to estimate the reindeer distribution. Since there are surplus zeros in the dataset, we also fit GLMM and mixed hurdle model with the data.

The size of the research area is 25 km^2 and it is located on the Storliden Mountain in north Sweden (see figure 1). The area contains marshes, forest, lakes and other kinds of landform. The reindeer graze freely from restraint from May to October in the place except a few days to gather them for calf-making. The data were collected from the year 2009 to 2012.

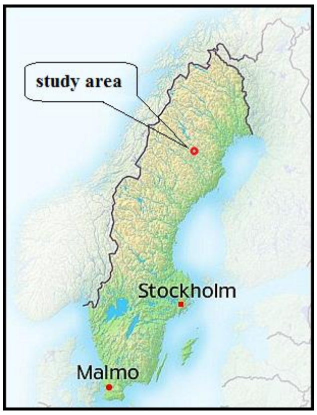

Figure 1 study area
Location in map of Sweden

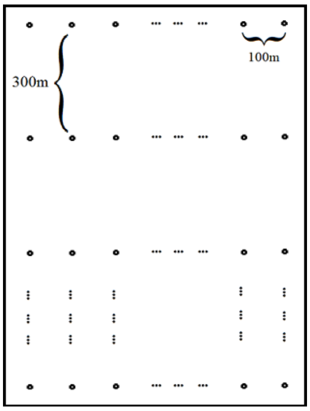

Figure 2 layout of
setting the transects and plots

In order to measure the reindeer distribution in the research region, the method faecal pellet-group counts was employed. Comparing with the other techniques, pellet-group counts is cheaper and simply that usually applied in wildlife investigation. Two ways were involved in calculate the pellet-group: faecal standing crop (FSC) and faecal accumulation rate (FAR). Transect survey design was utilized during the research. The study area was marked by different transects and the interval between each transects is 300 meter. There are distinct plots on each transect and all the plots center were tagged by a small wood stick. The radius for each plot is 2.18 meter. All the plots on each transect are 100 meters' apart from each other and each of them has an identity number (see figure 2). A pellet-group was counted if its center placed inside the plot. As the reindeer may move when they defecate, the pellets might split into some smaller groups. The separate group will be counted in such situation if it include more than 20 pellets.

Table 4 Pellet group summary

Pellet group	Counts	percentage
0	591	85.78%
1	79	11.47%
2	9	1.31%
3	7	1.02%
4	3	0.44%
total	**689**	**100.00%**

Treating the pellets group data which were collected in year 2010 and 2011 as response variable in our analysis. 355 plots are included in the study area. Because there are few missing values exist in the dataset, we delete these data which remains 689 counts to analysis. A summary of the response was illustrated in table 4. The maximum group number in a plot is 4. Only three observations contain 4 pellet groups. But 85.78% observations contains no pellet group which indicates that it is a typically zero inflated dataset.

We consider eight habitat variables as independent to begin with to fitting our model. There are six continuous explanatory variables and two categorical explanatory variables. The data of the forest age structure comes from the Swedish kNN-layer (ftp://salix.slu.se/download/skogskarta/). All the other geographical data were extracted from Lantmäteriet (http://www.lantmateriet.se).

Table 5 Continuous predictor summary

Number	Continuous variable	Range
1	Elevation	320 - 502 m amsl
2	Slope	0-18.75 degrees
3	Ruggedness index	$1.69*10^{-5}$ - $9.87*10^{-3}$
4	Distance to all roads	0-838 m
5	Distance to power lines	1598-4974 m
6	Forest age structure	0-104 years old

Table 6 Categorical predictor summary

Aspect classes	counts	percentage	Vegetation types	counts	percentage
Flat areas	17	2.47%	Broad-leaved forest	16	2.32%
Northwest slope	94	13.64%	Coniferous forest	187	27.14%
Northeast slope	239	34.69%	Clear cut	90	13.06%
Southeast slope	73	10.60%	Young forest	277	40.20%
Southwest slope	266	38.61%	Mire	44	6.39%
			Mixed forest	75	10.89%
total	**689**	**100.00%**	**total**	**689**	**100.00%**

There are totally 8 vegetation types in the dataset: broad-leaved forest, coniferous forest, clear cut, young forest, mire, mixed forest, house, mine and lake. Owing to no observations in the last 2 types, we skip them in the summary table.

4.2 Model fitting

GLM, GLMM and hurdle model are fitted separately with the following model specifications.

$$\begin{aligned} &\mu = E(y) \\ &\log(\mu) = Elevation + Slope + Ruggedness\_index + \log(dist\_road + 1) \\ &\qquad + \log(dist\_power + 1) + forest\_age + aspect\_class + vege\_type \end{aligned} \tag{23}$$

First we fit a GLM model as a starting point. As we mentioned in the former section, eight environmental variables were employed to estimate the reindeer spatial distribution. We start with a Poisson distribution for the response since they are counts. The canonical link log of Poisson distribution was applied (see Equation 23). The procedure was implemented by using the *glm* function in R software. We use stepwise selection method to select the predictors. Two predictors *clear cuts* and *Distance to power lines* (*PL-Distance*) are selected in the final model. We will take advantage of these two explanatory variables in following .

Treating *clear cuts* and *Distance to power lines* as fixed effects, we add the identity number of the locations as random effects to fit a GLMM model. The variable *plots_ID* in the parenthesis with a star stands for the random effects (see formula 24).

$$\begin{aligned} &\mu_{it} = E(y_{it}|u_i) \\ &\log(\mu_{it}) = \alpha + \beta_1 \log(dist\_power_{it} + 1) + \beta_2 Clear\_cuts_{it} + u_i \end{aligned} \tag{24}$$

Where the locations *i*=1, 2, …,*n*; the year *t* = 2010,2011; *α* is the intercept term; *β* represents coefficient of fixed effects, u_i is the random effect for location *i* with $\boldsymbol{u}= \{u_i\} \sim N(0,\Sigma_u)$.

At last, we fit mixed hurdle models with independent random effects and correlated random effects separately. For each response *y*, let P(*y*>0)=*p* and P(*y*=0)=1-*p*, a logistic regression model for p and a log-linear model for the mean μ of untruncated Poisson distribution can be represented as formula 25.

$$\begin{aligned} logit(p_{it}) &= \alpha_1 + \beta_{11}\log(dist\_power_{it}+1) + \beta_{12}Clear\_cuts_{it} + u_{1i} &&, \text{Bernoulli part} \\ \log(\mu_{it}) &= \alpha_2 + \beta_{21}\log(dist\_power_{it}+1) + \beta_{22}Clear\_cuts_{it} + u_{2i} &&, \text{T-Poisson part} \end{aligned} \quad (25)$$

Where the locations i=1, 2, …,n; the year t = 2010,2011; α_1 and α_2 are the intercept terms for Bernoulli part and truncated Poisson part; β_1. and β_2. represent the coefficient of fixed effects, u_{1i}, u_{2i} are the random effects for different parts with $u_{1i} \perp u_{2i}$ and $\boldsymbol{u_j} = \{u_{ji}\}_{i=1}^{n} = \sim N(0, \Sigma_j); j = 1,2$.

In order to fit the hurdle model with correlated random effects, we should define a neighborhood matrix first. It is produced by a neighborhood distance equals to 300 meters that which is the distance among near transects. The neighborhood matrix is generated by the codes in Appendix B.

4.3 Results

We conducted the new procedure to a real problem. In a mixed hurdle model, the fitted value for a single response can be calculated:

$$E(y) = (1-p)\times 0 + p\times \mu_t = (1-p)\times 0 + p\times \frac{\mu}{1-\exp(-\mu)} \quad (26)$$

Where p= P(y>0), μ_t is the expectation of truncated Poisson part, μ is the expectation of untruncated Poisson distribution in non-zero part.

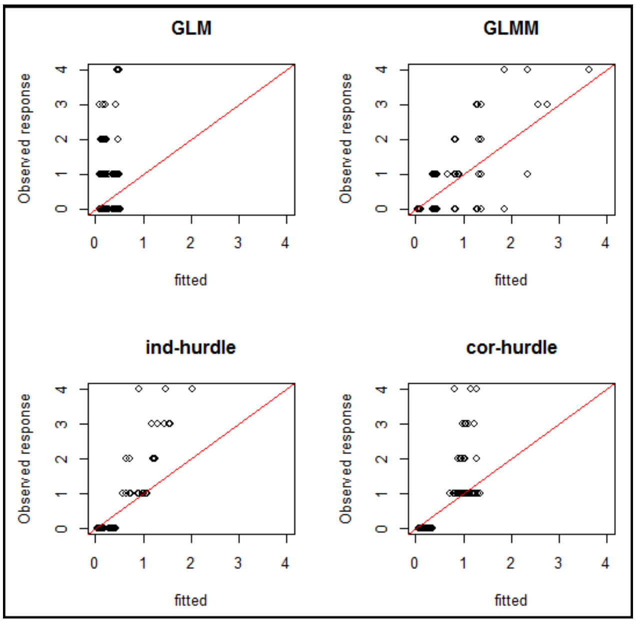

Figure 3 Plots of fitted vs. observed response for different models

We plot the fitted value and response of each model in figure 3. The summary statistics of all different models are displayed in Table 7.

If the estimation of the model fits the data perfectly, all the plots should on the line whose slope equals to 1 in Figure 3. In GLM model, the fitted values are quite small for all kinds of response which provide a worse fit than the others. It indicates that mixed model is more preferable for this dataset. Table 4 displays 85.78% observations are 0 and 11.47% observations equal to 1. Comparing the GLMM and hurdle model, we can see the GLMM model fits better for the data whose response larger than 2. However, it gives a wide range of estimate, especially for the response smaller than 2, while hurdle model fits better for most of observations. Comparing the profile likelihood in Table 7, it also shows the mixed models give a better fit than GLM model. The hurdle model with correlated random effects (-203.74) fits best among these models.

Table 7 Models summary

Variable	Estimates					
	GLM	GLMM	Ind-hurdle (mixed)		Cor-hurdle (mixed)	
			Bernoulli	T-Poisson	Bernoulli	T-Poisson
Intercept	-9.6997***	-9.1987**	-9.8226**	-6.6751	-8.9632**	-6.5857
	(3.1698)	(3.9537)	(4.2179)	(6.1194)	(4.1974)	(6.0528)
Clear cuts	0.8622***	0.9576***	1.1005***	0.3418	0.9168***	0.3245
	(0.2007)	(0.3165)	(0.3399)	(0.3586)	(0.3094)	(0.3495)
PL-Distance	2.2283***	1.8478*	2.1379*	1.6593	1.9846*	1.6306
	(0.8988)	(1.1259)	(1.2009)	(1.7296)	(1.1934)	(1.7098)
profile-likelihood	-360.15$^{\#}$	-236.56	-212.44		-203.74	

Remark: 1. Significant codes: 0.01 '***'; 0.05 '**'; 0.1 '*'. 2. $^{\#}$: This is the log-likelihood for GLM
3. The numbers in parenthesis are the corresponding standard error

Because the predictors all selected by stepwise method in R software, all the estimation are significant in GLM model. Table 7 presents that the fixed effects in GLMM are significant too. However, although the fixed effects in the Bernoulli parts are significant, but these are different in the Truncated Poisson part in mixed hurdle model. They are not coincide with each other. When we select the independent variables to fit a hurdle model, consider both the Bernoulli and truncated Poisson parts may give us a better fit. By looking at the estimation of correlated hurdle model, we may find that the reindeer are prefer to stay in the area with clear cuts. They do not like to live near the road which may indicates the reindeer do not want to disturbed by human beings.

5. Conclusion

Counts data with many zeros are encountered in many problems. Generalized linear models and its extensions are widely used to analyze non-normal data. But these methods may exhibit inadequate fitting and inference when encounter excessive zero counts. Hurdle model can be used to fit the zero-inflated data.

We provide a way to estimate the hurdle model with conditionally autoregressive random effects for the spatial data. In this paper:

1) We demonstrate that the iteratively reweighted least squares algorithm can be used to fit the mixed spatial hurdle model.

2) According to the extended algorithm, some modifications are made to the existing codes. A new procedure was implemented in R software to handle the hurdle model with correlated random effects within zero part and non-zero part.

3) A Monte-Carlo simulation study is conducted to understand the finite sample properties of the estimator. We employ the simulation study to independent and correlated random effects separately.

4) We apply the procedure to a real data problem to make a comparison with the other methods. It indicates that the new procedure can be used to fit the zero inflated data well.

In a word, we have extend an algorithm to estimate mixed hurdle model and implemented this numeric procedure in R. However, there are still some limitations. First, two types of link function (inverse and log) can be used to estimate the variance of random effects. The log link supplies a better estimate of the fixed effects parameter than the inverse link does. But the log link cannot estimate the two variance component parameters of a CAR model random effect. What's more, for some reasons such as calculating the exponential value of big numbers, the procedure sometimes runs slowly. Additional work is necessary to solve these problems in future research.

Bibliography

Alam, M., Rönnegård, L. & Shen, X., 2014. Fitting spatial models in the R package: hglm. *Working papers in transport, tourism, information technology and microdata analysis,* Issue 2014:01, pp. 1-8.

Hall, D. B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, pp. 1030-1039.

Jiang, J., 2007. *Linear and generalized linear mixed models and their applications.* New York: Springer.

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

Lee, Y. & Nelder, J. A., 1996. Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 619-678.

McCulloch, C. E. & Searle, S. R., 2001. *Generalized, Linear, and Mixed Models (1st ed.).* Canada: John Wiley & Sons, Inc.

Min, Y. & Agresti, A., 2005. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling,* pp. 1-19.

Moffatt, P. G., 2005. Hurdle Models of Loan Default. *The Journal of the Operational Research Society,* pp. 1063-1071.

Molas, M. & Lesaffre, E., 2010. Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine,* pp. 3294-3310.

Olsson, U., 2002. Generalized Linear Models: An Applied Approach. In: *Generalized Linear Models: An Applied Approach.* Lund: Studentlitteratur, pp. 31-44.

Pawitan, Y., 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. In: *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* London: Oxford University Press, pp. 435-469.

Skarin, A. & Rönnegård, L., 2011. Using kriging regression to detect change in reindeer distribution in relation to human development. *Procedia Environmental Sciences*.

Winkelmann, R., 2004. Health care reform and the number of doctor visits-an econometric analysis. *Journal of Applied Econometrics,* pp. 455-472.

Appendices

A: The procedure of hurdle model with spatial correlated random effects

```
# y: response variable
# x: design matrix of fixed effects
# z: vector of random effect (the procedure only consider one random effect)
#nb_matrix: neighborhood matrix
TP<- function(y,x,z,nb_matrix){
no.obs<- length(y)
# eigenvalue of neighborhood matrix
D.eigen<- eigen(nb_matrix)
D.eigenvalue<- as.numeric(D.eigen$values)
D.eigenvector<- as.matrix(D.eigen$vectors)

#create random effects design matrix
zf<-as.factor(z)
length.random<- nlevels(zf)
zz<-model.matrix(~0+zf)
tdim<-dim(zz)
attributes(zz)<-NULL
dim(zz)<-tdim
design_matrix.z<-zz

#create new design matrix for TP GLMM
z_new<- design_matrix.z%*%D.eigenvector
```

```
length.fix<- ncol(x)
x.new1<- cbind(x,z_new)
zero.matrix<- matrix(0,nrow=length.random,ncol=length.fix)
x.new2<- cbind(zero.matrix,diag(length.random))
x.new<- rbind(x.new1,x.new2)
length.par<- length.random+length.fix

#Initial guess
reg<- lm(y~x-1)
beta_m<- as.numeric(reg$coef)
b<- rep(0,length=length.random)
th1<- 0.7
th0<- abs(min(th1*D.eigenvalue))+0.8
w0<-as.numeric(th0+th1*D.eigenvalue)

###loop
Iter0<- TRUE
i<- 1
while (Iter0){

# inside loop
Iter<- TRUE
j<- 1
while (Iter){

# mu=E(y)=d(mu)/d(eta) They are same only for the poisson model
mu<- as.vector(exp(x%*%beta_m+z_new%*%b))

s<-y-mu-mu/(exp(mu)-1)
w<-mu+(exp(mu)-mu*exp(mu)-1)*mu/(exp(mu)-1)^2
w.newvalue<- c(w,as.numeric(w0))
x_new<- x.new*sqrt(w.newvalue)
sai<- x%*%beta_m+z_new%*%b+s/w
sai.n<- sai*sqrt(w)
sai.n<- c(sai.n,rep(0,length.random))
m1<- lm(sai.n~x_new-1)
fc1<- as.numeric(m1$coefficients)
j<- j+1
if (max(abs(c(beta_m,b)-fc1))<1e-4) Iter=FALSE
beta_m<- fc1[1:length.fix]
b<- fc1[(length.fix+1):length.par]

if (j>100){
Iter=FALSE
print("Max interations. Not converged")
}
}

b2<-b^2
leng.hat<-length(hatvalues(m1))
leng.b<-length(b)
hv1<- hatvalues(m1)[(leng.hat-leng.b+1):leng.hat]
wt2<- (1-hv1)/2

##############################
#       different link selection          #
##############################
# inverse link start----------
#car.model<- glm(I(b2/(1-hv1))~D.eigenvalue,family=Gamma(link=inverse),
#weights=wt2,start=c(.6,.01))
#w0<-predict(car.model)
#i<- i+1
#Conv<-max(abs(c(th0,th1)-coef(car.model)))
#if (Conv<1e-4) Iter0=FALSE
#th0<- car.model$coefficients[1]
#th1<- car.model$coefficients[2]
#inverse link end-------------
# log link start----------------
car.model<- gam(I(b2/(1-hv1))~s(D.eigenvalue),family=Gamma(link=log),weights=wt2)
th<-as.numeric(car.model$coefficients)
if(i>1){
Conv<-max(abs(th0-th))
if (Conv<1e-4) Iter0=FALSE
}
i<- i+1
th0<-th
w0<- 1/as.numeric(car.model$fitted.values)
#log link end-------------------
if (i>50){
Iter0=FALSE
}
}

print(round(data.frame(summary(m1)$coef[1:length.fix,]),4))
cat('\n')
print(round(summary(car.model)$coef,4))

}
```

B: The procedure to generate the neighborhood matrix

```
# x and y are the coordinate of spatial locations,
# if the distance between two locations i and j
#less than dist, then D[i,j]=1, otherwise D[i,j]=0
# and all the D[i,i]=0
nb_matrix<-function(x,y,dist){

# create neighborhood matrix
D<-NULL

# loop
i<-1
while(i<=length(x)){
j<-1
line<-NULL
while (j<=length(y)){
if(i==j) {
line<-c(line,0)
}
else if((x[i]-x[j])^2+(y[i]-y[j])^2<dist^2) {
line<-c(line,1)
}
else{
```

```
line<-c(line,0)
}
j<-j+1
}
D<-cbind(D,line)
i<-i+1
}
return(D)
}
```

C: The procedure of hurdle model with independent random effects

```
# y: response variable
# x: design matrix of fixed effects
# z: vector of random effect (the procedure only consider one random effect)
TP.uncor<- function(y,x,z){
no.obs<-length(y)

#create random effects design matrix
zf<-as.factor(z)
length.random<- nlevels(zf)
zz<-model.matrix(~0+zf)
tdim<-dim(zz)
attributes(zz)<-NULL
dim(zz)<-tdim
design.z<-zz

#create new design matrix for GLMM
length.fix<- ncol(x)
x.new1<- cbind(x,design.z)
zero.matrix<- matrix(0,nrow=length.random,ncol=length.fix)
x.new2<- cbind(zero.matrix,diag(length.random))
x.new<- rbind(x.new1,x.new2)
length.par<- ncol(x.new)

#Initial guess
reg<- lm(y~x-1)
beta<- reg$coef
b<- rep(0,length=length.random)
th0<- 0.5
w0<- rep(1/th0,length.random)

###loop
Iter0<- TRUE
i<- 1
while (Iter0){

# inside loop
Iter<- TRUE
j<- 1
while (Iter){

# mu=E(y)=d(mu)/d(eta)    they are same only for the poisson model
mu<- as.vector(exp(x%*%beta+design.z%*%b))

#gradient
s<- y-mu-mu/(exp(mu)-1)

#Hessian new
w<- mu+(exp(mu)-mu*exp(mu)-1)*mu/(exp(mu)-1)^2
w.newvalue<- c(w,w0)
w.new<- diag(w.newvalue)

#fit the linear model to estimate the parameter
x_new<- x.new*sqrt(w.newvalue)
sai<- x%*%beta+design.z%*%b+s/w
sai.n<- sai*sqrt(w)
sai.n<- c(sai.n,rep(0,length.random))
m1<- lm(sai.n~x_new-1)
fc1<- as.numeric(m1$coefficients)

j<- j+1
if (max(abs(c(beta,b)-fc1))<1e-5) Iter=FALSE
beta<- fc1[1:length.fix]
b<- fc1[(length.fix+1):length.par]
if (j>50){
Iter=FALSE
print("Max interations. Not converged")
}
}

b2<-b^2
leng.hat<-length(hatvalues(m1))
leng.b<-length(b)
hv1<- hatvalues(m1)[(leng.hat-leng.b+1):leng.hat]
wt2<- (1-hv1)/2
car.model<- glm(I(b2/(1-hv1))~1,family=Gamma(link = "log"),weights=wt2,start=0.2)
w0<-predict(car.model,type="response")
w0<-1/w0
i<- i+1

if (abs(th0-coef(car.model))<1e-5) Iter0=FALSE
th0<- car.model$coefficients[1]
if (i>100){
Iter0=FALSE
}
}

print(round(data.frame(summary(m1)$coef[1:length.fix,]),4))
cat('\n')
print(round(summary(car.model)$coef,4))

}
```

D: Simulation of hurdle model with correlated random effects

```
# times : how many times of simulate
# n : number of locations in the generated spatial dataset
# k : number of observation in each location
# alpha    : the intercept of the fixed effects
# beta1, beta2 : the parameters of the fixed effects
# tau, rou : the parameters of the car model
```

```
simulate<-function(times,n,k,alpha,beta1,beta2,tau,rou){
set.seed(1234)
library(aster)
library(MASS)
source("C:\\Users\\Qiang\\Documents\\R \\neibor-matrix.r")
source("C:\\Users\\Qiang\\Documents\\R \\hurdle-cor.r")

beta<-c(alpha,beta1,beta2)
th0<-1/tau
th1<--rou/tau
#---create neighborhood matrix---
sq.n<-sqrt(n)
d1<- runif(n,0,sq.n)
d2<- runif(n,0,sq.n)
Dm<- nb_matrix(d1,d2,sq.n/(2*sqrt(pi)))
var.b<- tau*(solve(diag(n)-rou*Dm))

#---create design matrix of random---
de.z<-rep(1:n,each=k)
de.zf<-as.factor(de.z)
design.z<-model.matrix(~0+de.zf)
tdim<-dim(design.z)
attributes(design.z)<-NULL
dim(design.z)<-tdim

Iter<- TRUE
t<- 1
ParToSave<-matrix(NA,nrow=times,ncol=4)
while(t<=times){
#-----------------------------------------------

#----fixed part-------
x1<- runif(n*k,0,1)
x2<- runif(n*k,0,2)
x<- cbind(1,x1,x2)

#----random part------
b<- as.numeric(mvrnorm(1,rep(0,n),var.b))

#----response part----
eta<- x%*%beta+design.z%*%b
mu<- exp(eta)
y<- rnzp(n*k, mu, xpred = 1)
#----------------------------------------------

aa<- try(TP(y,x,de.z,Dm))
if(!(inherits(aa,"try-error"))){
if(aa[4]<50){
ParToSave[t,]<-as.numeric(aa[1:4])
}
}
t<- t+1

}
return(ParToSave)
}
```

E: Simulation of hurdle model with independent random effects

```
# times : how many times of simulate
# n : number of locations in the generated spatial dataset
# k : number of observation in each location
# alpha : the intercept of the fixed effects
# beta1, beta2 : the parameters of the fixed effects
# tau, rou : the parameters of the car model
simulate.un<-function(times,n,k,alpha,beta1,beta2,tau){
set.seed(1234)
library(aster)
library(MASS)
source("C:\\Users\\Qiang\\Documents\\R \\hurdle-ind.r")

beta<-c(alpha,beta1,beta2)

#---create design matrix of random---
de.z<-rep(1:n,each=k)
de.zf<-as.factor(de.z)
design.z<-model.matrix(~0+de.zf)
tdim<-dim(design.z)
attributes(design.z)<-NULL
dim(design.z)<-tdim

t<- 1
ParToSave<-matrix(NA,nrow=times,ncol=4)
while(t<=times){

#----fixed part-------
x1<- runif(n*k,0,1)
x2<- runif(n*k,0,2)
x<- cbind(1,x1,x2)

#----random part------
b<- as.numeric(rnorm(n,0,sd=sqrt(tau)))

#----response part----
eta<- x%*%beta+design.z%*%b
mu<- exp(eta)
y<- rnzp(n*k, mu, xpred = 1)

aa<- try(TP.uncor(y,x,de.z))
if(!(inherits(aa,"try-error"))){
if(aa[5]<50){
ParToSave[t,]<-as.numeric(aa[1:4])
}
}
t<- t+1
}
return(ParToSave)
}
```

Acknowledgement

I would like to express the appreciation to my supervisor Md. Moudud Alam. He helped me to check my codes and gave me some suggestions. His attitude to do research gave me a deep impression. Without his supervision and constant help this dissertation would not have been possible. I would also want to say thank you to Professor Lars Rönnegård who introduced me of the original dataset and lend me the related books.

In addition, I would like to say thanks to the teachers in Dalarna University who gave me lectures. Thanks for your teaching, and thanks to Dalarna University. I did have a happy time here.