Generalized Estimating Equations for Zero-Inflated Spatial Count Data[1]

Anthea Monod
Department of Mathematics, Swiss Federal Institute of Technology (EPFL),
Station 8, CH-1015 Lausanne, Switzerland; anthea.monod@epfl.ch
Under the supervision of Prof. Stephan Morgenthaler

Abstract: We consolidate the zero-inflated Poisson model for count data with excess zeros (Lambert, 1992) and the two-component model approach for serial correlation among repeated observations (Dobbie and Welsh, 2001) for spatial count data. This concurrently addresses the problem of overdispersion and distinguishes zeros that arise due to random sampling from those that arise due to inherent characteristics of the data. We give a general quasi-likelihood and derive corresponding score equations for the zero-inflated Poisson generalized linear model. To introduce dependence, a spatial-temporal correlation structure comprising forms for fixed time, fixed location, and neighbor interactions is required; construction using techniques from the theory of Markov point processes is investigated.

Keywords: Generalized estimating equation (GEE), spatial count data, zero-inflated counts, zero-inflated Poisson model, nearest-neighbor marked Markov point processes, Dirichlet tessellation.

1 Introduction

Let y_{it} denote the number of occurrences of an event observed at $t = 1, \ldots, T_i$ time points for each subject $i = 1, \ldots, n$, and let $\mathbf{x}_{it} \in \mathbf{R}^q$ be a vector of measured covariates. Such data is often modeled through a generalized linear model to provide greater flexibility, specifying a form for the expectation, $E[Y_{it}] = \lambda_{it} = g^{-1}\big(\mathbf{x}'_{it}\boldsymbol{\beta}\big)$, with $\boldsymbol{\beta}$ a $q \times 1$ vector of unknown parameters, and the link function $g(\cdot)$ commonly taken to be the log function. Under a Poisson distribution, the variance is equal to the mean, $\mathrm{Var}(Y_{it}) = \lambda_{it} = E[Y_{it}]$, which in practice may be too restrictive; often the data exhibit $E[Y_{it}] = \mathrm{Var}(Y_{it})$, known as overdispersion.

Lambert (1992) has presented zero-inflated Poisson (ZIP) regression, giving rise to a new class of regression models for count data with an abundance of zero observations. In a ZIP model, the non-negative integer response Y is assumed to be distributed as a mixture of a Poisson distribution with parameter λ_{it}, and a distribution with point mass of one at the value zero, with mixing probability α_{it}; the non-zeros and a portion of the zeros are modeled by the usual Poisson probability.

[1] Research supported in part by the Swiss National Science Foundation, Grant No. FN 200021-116146

Dobbie and Welsh (2001) adapt the generalized estimating equations approach of Liang and Zeger (1986) to zero-inflated spatial count data, addressing dependence by incorporating a correlation matrix. They model the abundance of zeros via a two-component approach: the zeros are modeled separately from the non-zeros; first, absence versus presence (zero versus non-zero) is described by a logistic model, and then conditional on presence, the non-zero counts are described by a truncated Poisson distribution.

We work in the context of a Poisson generalized linear model, consolidating the two aforementioned approaches to construct generalized estimating equations for the zero-inflated Poisson generalized linear model comprising spatial-temporal dependence. Attributing some of the zeros to the Poisson distribution avoids conditioning on the responses, and provides a more intuitive approach to occurrence of zeros in the data. The data of interest are weekly counts of Noisy Friarbirds (*Philemon corniculatus*) recorded by observers for the Canberra Garden Bird Survey: attributing a probability weight of zero observations to a point mass distribution and its complement to a Poisson distribution allows for the distinction between zero counts arising due to an inherent characteristics that may induce zero observations (*e.g.* inadequacy of the region where measurements were taken for the survival or reproduction of Noisy Friarbirds), and zero counts arising at random. In considering dependence, the theory of nearest-neighbor Markov point processes proves to be useful in constructing covariance forms for the zero-inflated spatial data.

In this paper, we detail the theoretical results behind the work to be presented at the 2011 European Regional Conference of The International Environmetrics Society (TIES), "Spatial Data Methods for Environmental and Ecological Processes – 2nd Edition".

2 Methodology

We implement the zero-inflated Poisson model of Lambert (1992) to address overdispersion, and obtain a likelihood and score equations, which, following Dobbie and Welsh (2001), turn out to be generalized estimating equations in the style of Liang and Zeger (1986); we incorporate dependence into the model following Diggle *et al.* (2009). In constructing a space-time dependence structure, we focus on the neighbor interaction component and outline the theory of Markov point processes relevant to this aspect.

The Zero-Inflated Poisson Generalized Linear Model. A non-negative, integer-valued random variable describing a discrete number of occurrences for a cross-sectional unit i at time period t is said to follow a *zero-inflated Poisson distribution* with parameter $\lambda_{it} \in (0, \infty)$ and mixing probability $\alpha_{it} \in (0, 1)$ if

$$Y_{it} \sim \begin{cases} 0 & \text{with probability } \alpha_{it}, \\ \text{Poisson}(\lambda_{it}) & \text{with probability } (1 - \alpha_{it}). \end{cases} \tag{1}$$

The parameters $\lambda_{it}, \alpha_{it}$ are allowed to depend on auxiliary covariate information, for simplicity we assume the same auxiliary information. It follows that $E[Y_{it}] = (1-\alpha_{it})\lambda_{it}$ and $\mathrm{Var}(Y_{it}) = (1-\alpha_{it})\lambda_{it}(1+\alpha_{it}\lambda_{it})$ and indeed $E[Y_{it}] < \mathrm{Var}(Y_{it})$.

Under this model, the observations are generated by

$$\mathrm{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}) = \alpha_{it}\mathbb{1}(y_{it}=0) + (1-\alpha_{it})\frac{\exp\left\{y_{it}\mathbf{x}'_{it}\boldsymbol{\beta} - e^{\mathbf{x}'_{it}\boldsymbol{\beta}}\right\}}{y_{it}!}, \tag{2}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function; the probability of observing a zero is $\mathrm{Prob}(Y_{it} = 0|\mathbf{x}_{it}) = \alpha_{it} + (1-\alpha_{it})\exp\left\{-e^{\mathbf{x}'_{it}\boldsymbol{\beta}}\right\}$.

Likelihood and Score Equations. The log-likelihood for the zero-inflated Poisson model is $\ell(\alpha_{it}, \boldsymbol{\beta}) = \sum_{i,t:y_{it}=0} \log\left(\alpha_{it} + (1-\alpha_{it})e^{-e^{\mathbf{x}'_{it}\boldsymbol{\beta}}}\right) + \sum_{i,t:y_{it}>0}\left(y_{it}\mathbf{x}'_{it}\boldsymbol{\beta} - e^{\mathbf{x}'_{it}\boldsymbol{\beta}} - \log y_{it}!\right)$, which gives the following score equation with regard to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\alpha_{it}, \boldsymbol{\beta}) = \sum_{i,t:y_{it}=0}(y_{it}-\lambda_{it})\frac{\mathrm{Prob}(Y_{it}=0) - \alpha_{it}}{\mathrm{Prob}(Y_{it}=0)} + \sum_{i,t:y_{it}>0}(y_{it}-\lambda_{it})\mathbf{x}_{it} = 0. \tag{3}$$

Modeling the mixing probability α_{it} as any differentiable function of another parameter γ, $\alpha_{it} = \alpha_{it}(\gamma)$, the score equation for the ZIP model with regard to γ is

$$\frac{\partial}{\partial\gamma}\ell(\alpha_{it}, \boldsymbol{\beta}) = \sum_{i,t}\frac{\mathrm{Prob}(Y_{it}>0)}{\mathrm{Prob}(Y_{it}=0)}\frac{\partial\alpha_{it}}{\partial\gamma}\frac{1}{1-\alpha_{it}} = 0. \tag{4}$$

Note that ratio of probabilities in this latter equation provides an intuitive odds-ratio interpretation of the weighting between the two probability components.

Introducing Dependence. Following Dobbie and Welsh (2001) and Diggle *et al.* (2009) in the setting of marginal models, we introduce dependence by extending the score equations (3) and (4) to comprise a $2T_i \times 2T_i$ spatial variance-covariance matrix. Diggle *et al.* (2009) show that for marginal models under appropriate parameterizations, the score equations assume a form of a generalized estimating equation (Liang and Zeger, 1986), whose solution gives a consistent estimator:

$$\left(\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right)'\mathrm{Var}(Y)^{-1}(Y-\mu) = 0. \tag{5}$$

In the spatial-temporal setting, the covariance requires structures for fixed time, fixed location, and neighboring interactions. Models for the former cases are readily available in time series analysis and spatial statistics literature. In our application to Noisy Friarbird counts, the latter case is of particular interest, since, depending on the region partition, observations in one region is likely to influence that in nearby regions: vicinities of unsuitable habitat regions may also be less suitable, thus influencing a low-valued observation. This motivates the use of techniques of nearest-neighbor Markov point processes and random tessellations to address neighbor interaction as well as region partitioning.

Models for observations generated by marked Markov point processes can be augmented to allow interactions to depend on the realization of the process by generalizing the spatial Markov property, as shown by Baddeley and Møller (1989). Moreover, the spatial interaction in a marked Markov point process can be analyzed conditional on the positions of the points, since the conditional distribution of the marks given the point configuration is a Gibbs process on the finite graph defined by the points. Dirichlet tessellation is shown to satisfy nearest-neighbor conditions in the construction of such processes where each point interacts with its neighbors, notably that of the invariance of connectivity between any two points under the addition of a new point, unless it is a neighbor of both points.

3 Concluding Remarks

The spatial-temporal zero-inflated Poisson generalized linear model addresses overdispersion present in space-time data comprising excess zeros, while providing greater flexibility in the modeling and interpretation of zeros due to random sampling and those due to characteristics of the data, and may be extended to incorporate spatial-temporal dependence. The nature of such data motivates the consideration of neighboring interactions when constructing forms for dependence, which then inspires the use of techniques of nearest-neighbor Markov point processes, allowing for the generation of spatial points with interaction that is conditional on their positions. This two-fold approach to the challenges of zero-inflated, correlated spatial-temporal data will indeed prove to be applicable in various ecological and biological contexts, and useful in general applications in diverse fields of science.

References

Baddeley A., Møller J. (1989) Nearest-Neighbour Markov Point Processes and Random Sets, *International Statistical Review*, 57(2), 89–121.

Diggle P. J., Heagerty P., Liang K.-Y., Zeger S. (2009) *Analysis of Longitudinal Data* 2nd ed., Oxford University Press, Oxford.

Dobbie M. J., Welsh A. H. (2001) Modelling Correlated Zero-Inflated Count Data, *Aust. N. Z. Stat.*, 43(4), 431–444.

Lambert D. (1992) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34(1), 1–14.

Liang K.-Y., Zeger S. (1986) Longitudinal Data Analysis Using Generalized Linear models, *Biometrika*, 73(1), 13–22.