Space-Time Zero-Inflated Count Models of Harbor Seals

Jay M. Ver Hoef[1] and John Jansen[1]

[1] National Marine Mammal Lab,, 7600 Sand Point Way NE, Bldg 4, Seattle, WA 98115-6349, Voice: (206) 526-4025, FAX: (206) 526-6615, E-mail: jay.verhoef@noaa.gov

Abstract: Environmental data are spatial, temporal, and often come with many zeros. In this paper, we take the standard formulation of a zero-inflated Poisson (ZIP) model, as well as an alternative parameterization, and develop a space-time model to investigate haulout patterns of harbor seals on glacial ice. The data consist of counts, for 18 dates on a lattice grid of samples, of harbor seals hauled out on glacial ice in Disenchantment Bay, a coastal bay near Yakutat, Alaska. A space-time ZIP was constructed by using spatial conditional autoregressive model (CAR) model and a temporal first-order autoregressive model (AR1) as random effects in ZIP regression model. Because seals are unlikely to be undetected, we consider another model that completely specifies and separates the binary from the count process, but still has an inflated number of zeros. We compare this model to the standard ZIP. Both models indicate that ice density plays a strong role in where seals haul out, with highest haulout probabilities and counts at intermediate ice densities. We create maps of smoothed prediction rates for harbor seal haulouts based on ice density and other covariates.

Keywords: autoregressive; AR1; CAR.

1 Introduction

Environmental data are spatial, temporal, and often come with many zeros. Statisticians are developing models of increasing complexity to handle these data. Time series (e.g., Brockwell and Davis, 1991), spatial statistics (e.g., Cressie, 1993), and zero-inflated Poisson (ZIP) regression (e.g., Lambert, 1989) are all well-developed subjects. There are increasing numbers of examples where models combine these subjects, such as space-time models for Gaussian data (e.g. Wikle et al., 1998) and spatial ZIP models (e.g. Agarwal et al., 2002). Wikle and Anderson (2003) developed a space-time ZIP model for tornado counts that is very similar to our development.

2 Data

The data consist of counts of harbor seals hauled out on glacial ice in Disenchantment Bay, a coastal bay near Yakutat, Alaska. Aerial surveys

were conducted twice weekly, weather permitting, starting 27 May and ending on 4 August in 2002. Surveys were flown between 13:00 - 15:00 h (ADT) to coincide with the daily peak in numbers of seals hauling out. A single engine aircraft (Cessna 206; Yakutat Coastal Airways Inc., Yakutat, AK) was flown at a target speed of 90-100 knots and altitude of 305 m (1000 ft). There are 18 time events that we index $i = 1, 2, \ldots, 18$. The ice and seal point data were summarized into a lattice of 400 $\times$ 400 m cells for the entire study area; the spatial locations are on a grid that we will index arbitrarily, $j = 1, 2, \ldots, m$. Grid cells that did not have ice had no possibility for seals to haul out, and not all grid cells had ice for each date. In all there were 2489 cells that contained ice over the 18 time periods.

3 Models for Zero-Inflated Count Data

A space-time ZIP can be constructed by using spatial conditional autoregressive model (CAR) model and a temporal first-order autoregressive model (AR1) as random effects in a ZIP model. A ZIP regression model is given by

$$Z_{i,j}|Y_{i,j} = \begin{cases} 0 & \text{if } Y_{i,j} = 0, \\ \text{Poi}(\lambda_{i,j}) & \text{if } Y_{i,j} = 1. \end{cases} \tag{1}$$

where $\text{Poi}(\lambda_{i,j})$ is a Poisson distribution with mean function $\lambda_{i,j}$ and $Y_{i,j}$ has a Bernoulli distribition with mean function $p_{i,j}$; $Y_{i,j} \sim \text{Bern}(p_{i,j})$, for the ith time and the jth spatial location. Now we use link functions, as is common for generalized linear models (McCullough and Nelder, 1989) to relate the means of these distributions to a linear mixed model,

$$\begin{aligned} \log(\lambda_{i,j}) &= \nu_i + \mathbf{x}'_{i,j}\boldsymbol{\beta} + \epsilon_{i,j}, \\ \text{logit}(p_{i,j}) &= \mu_i + \mathbf{x}'_{i,j}\boldsymbol{\alpha} + \delta_{i,j}, \end{aligned} \tag{2}$$

where $\text{logit}(a) \equiv \log\left(\frac{a}{1-a}\right)$ and $\mathbf{x}_{i,j}$ are covariates that vary both spatially and temporally. In our case, variables such as percent ice in a sample unit changes spatially and temporally due to weather and currents between observations. In a fixed effects model, we would assume that ν_i and μ_i are separate means for each time; here we treat them as separate linear models for covariates that only vary temporally, such as the weather on the day of the photograph that affects all spatial locations equally.

$$\begin{aligned} \nu_i &= \nu_0 + \mathbf{t}'_i\boldsymbol{\eta} + \xi_i, \\ \mu_i &= \mu_0 + \mathbf{t}'_i\boldsymbol{\gamma} + \tau_i, \end{aligned} \tag{3}$$

where $\mathbf{t}_i$ are time-varying covariates. It is here that we allow temporally autocorrelated errors, which are modeled with AR1 models,

$$\begin{aligned} \xi_i &= \phi_\xi \xi_{i-1} + \sigma_\xi W_{\xi,i}; \quad i > 1, \\ \tau_i &= \phi_\tau \tau_{i-1} + \sigma_\tau W_{\tau,i}; \quad i > 1, \end{aligned} \tag{4}$$

where $W_{\xi,i}$ and $W_{\tau,i}$ are independent Gaussian random variables. In (2), we assume that each time period has a separate and independent realization of a spatial process for $\epsilon_{i,j}$ and $\delta_{i,j}$. We use a spatial conditional autoregressive model (CAR) (see Besag 1974 and Cressie, 1993) for each time period, but allow the autocorrelation parameters to be common across time periods. Hence,

$$\begin{aligned} \boldsymbol{\delta}_i &= \text{Gau}(\mathbf{0}, \sigma^2_\delta(\mathbf{I} - \rho_\delta \mathbf{C})^{-1}\mathbf{M}) \\ \boldsymbol{\epsilon}_i &= \text{Gau}(\mathbf{0}, \sigma^2_\epsilon(\mathbf{I} - \rho_\epsilon \mathbf{C})^{-1}\mathbf{M}) \end{aligned} \tag{5}$$

where the spatial process for the ith time period $\boldsymbol{\delta}_i$ is independent of the spatial process $\boldsymbol{\delta}_{i'}$ when $i \neq i'$, and similarly the spatial process $\boldsymbol{\epsilon}_i$ is independent of the spatial process $\boldsymbol{\epsilon}_{i'}$ when $i \neq i'$. Gau$(\cdot, \cdot)$ is a (multivariate) Gaussian (normal) distribution. We defined a neighbor of a sample as any other sample with its centroid within 1 km. The weights in $\mathbf{C}$ were row-standardized (Haining, 1990, pg. 82); that is, each row in $\mathbf{C}$ contains all zeros except for columns that indicate a neighbor, and these values are the inverse of the number of neighbors for that sample. The matrix $\mathbf{M}$ is a diagonal matrix where the diagonal elements contain the inverse of the number of neighbors.

3.1 A Nonmixture Model

The ZIP model is important when zeros are a mixture of two processes; a binary process and a count process that includes zero whenever the binary process has a value of one. When considering harbor seal counts on ice, as in our application, the binary process is the absence or presence of harbor seals, and the count process is the number of seals. If there are detectability issues, this model is appropriate because it expresses the idea that an observed count can be 0 even though seals are present; i.e., some seals are undetected. However, in our application, we have aerial photographs of very high resolution, and seals are unlikely to be undetected. Hence, we consider a model that completely specifies and separates the binary from the count process. They are no longer a mixture, but there is clearly an overabundance of zeros in comparison to a simple Poisson distribution, so it is logically tied to ZIP models and can be compared purely on a model-fitting basis. We term this the "Poisson+1/Binary" model and denote it P1B. For it's formulation as a space-time model, we modify (1) to be,

$$Z_{i,j}|Y_{i,j} = \begin{cases} 0 & \text{if } Y_{i,j} = 0, \\ \text{Poi}(\lambda_{i,j}) + 1 & \text{if } Y_{i,j} = 1. \end{cases} \tag{6}$$

The rest of the model follows exactly as in the ZIP, using (2-5).

3.2 Priors

We put diffuse priors on all regression parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$. Because these are modeled on a log scale, there are computational instabilities if they are

allowed to get too large, so we let each regression parameter have a normally distributed prior with a variance of 10. The autoregression parameters for both space, (ρ_δ and ρ_ϵ) and time (ϕ_τ and ϕ_ξ) are bounded from -1 to 1, but we did not expect any negative autocorrelation, so we used uniform priors from 0 to 1. For the variance parameters of the random effects (σ_δ^2, σ_ϵ^2), we let the square root be uniformly distributed between 0 and 10; again, to keep the random effects from becoming too large and causing numerical instability.

4 Results

We present results on estimated space-time autocorrelation parameters, estimated regression coefficients, and smoothed prediction maps.

References

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

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

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods, 2nd Ed.* New York: Springer-Verlag.

Cressie, N. (1993). *Statistics for Spatial Data, Revised Edition.* New York: John Wiley and Sons.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

Lambert, D. (1989). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

McCullough, P. and Nelder, J.A. (1989). *Generalized linear models, 2nd Ed.*. New York: Chapman and Hall.

Wikle, C.K., and Anderson, C.J. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *Journal of Geophysical Research*, **108**, No. D24, 9005, doi:10.1029/2002JD0028006.

Wikle, C.K., Berliner, L.M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117-154.