Environmental and Ecological Statistics **9,** 341–355, 2002

Zero-inflated models with application to spatial count data

DEEPAK K. AGARWAL,[1] ALAN E. GELFAND[2] and STEVEN CITRON-POUSTY[3]

[1] *AT&T Shannon Research Labs, Florham Park, NJ 07932-0971*
[2] *Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251*
[3] *Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269*

Received October 2000; Revised October 2001

Count data arises in many contexts. Here our concern is with spatial count data which exhibit an excessive number of zeros. Using the class of zero-inflated count models provides a flexible way to address this problem. Available covariate information suggests formulation of such modeling within a regression framework. We employ zero-inflated Poisson regression models. Spatial association is introduced through suitable random effects yielding a hierarchical model. We propose fitting this model within a Bayesian framework considering issues of posterior propriety, informative prior specification and well-behaved simulation based model fitting. Finally, we illustrate the model fitting with a data set involving counts of isopod nest burrows for 1649 pixels over a portion of the Negev desert in Israel.

Keywords: conditionally autoregressive prior, Langevin diffusions, latent variables, posterior propriety

1. Introduction

Depending on the context, count data may be modeled in a variety of ways, e.g., using Poisson, negative binomial, binomial, beta-binomial or hypergeometric distributions. These familiar parametric families are limited in shape and tail behavior and so if, after adjusting for covariates, excess heterogeneity still persists, random effects are often introduced. For spatial count data, Poisson regression is most frequently used and the random effects are introduced using spatial models. For instance, conditionally autoregressive (CAR) priors are typically used with lattice or areal unit data (e.g., Clayton and Bernardinelli, 1991), spatial processes are used with geostatistical or point-referenced data (Diggle *et al*., 1999).

1352-8505 © 2002 Kluwer Academic Publishers

Here we treat a special problem with such data, the case of an excessive number of zero counts in the data. Nonspatial examples are common in industrial applications where a reliable manufacturing process moves back and forth between a perfect state in which defects are extremely rare and an imperfect state in which number of defects follow say a Poisson distribution. (e.g., Lambert, 1992; Ghosh *et al.*, 1998). The spatial problem that concerns us involves counts of nest burrows in each of more than 1600 pixels; 82% of these counts are 0. However, one can similarly envision, for example, species count data where, at most of the sampling sites, the species was not observed. A natural way to model such data is to put a point mass p at 0. That is, with probability p, we sample a degenerate distribution at 0 and with probability $(1-p)$ we sample say a Poisson (λ) distribution. Such models are called zero inflated Poisson (ZIP) models in the literature. Cohen (1963) and Johnson and Kotz (1969) discuss ZIP models without covariates. Heilbron (1989) studies ''zero-altered'' Poisson and negative binomial regression models in the context of high risk behavior in gay men. Lambert (1992) employs the ZIP model in a regression set up defining linear regressions for both λ and p. Her approach uses the E-M algorithm (Dempster *et al.*, 1977) to obtain the maximum likelihood estimate (MLE). The solution to the likelihood equations need not be unique, the E-M algorithm need not converge. Interval estimates rely on customary likelihood asymptotics; they are based on normal approximations which require the log-likelihood surface to be approximately quadratic near the MLE. Ghosh *et al.* (1998) discuss fully Bayesian methods for fitting ZIP models, obtaining exact credible intervals.

In this paper, we discuss the use of Bayesian methods to analyze zero contaminated count data that, in the presence of covariate information, are correlated in space. After examining general parametric classes of zero inflated models obtained by adding a point mass at 0, we focus primarily on the Poisson case. We formulate regression specifications for both p and λ incorporating spatial random effects in the latter. In particular, $\log(\lambda)$ is assumed to be a linear function of a set of covariates and a spatial random effect while log $(p/1-p)$ is assumed to be a linear function of a set of covariates only. We refer to such models as spatial ZIP regressions.

We also note that an ordinary Poisson regression with small λ's for the pixels but very strongly correlated random effects (essentially $p = 0$ in the foregoing modeling) need not provide a sufficiently flexible model. In usual Poisson regression, posterior propriety under improper priors on the regression coefficients is discussed in, for example, Diaconis and Ylvisaker (1979) and Laud and Ibrahim (1991). The ZIP case introduces an additional wrinkle which we discuss. One solution is to use proper and, in fact, fairly informative priors (to obtain well behaved simulation-based model fitting). This leads us to devise techniques for proper prior specification for the regression parameters in the p-regression. Posterior sampling for the parameters in the λ-regression is implemented within a Gibbs sampler using a Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996). We illustrate our method by fitting the proposed model to the previously mentioned dataset which records isopod nest burrow counts for pixel areal units.

In Section 2, we develop zero-inflated modeling details. In Section 3, issues of posterior propriety are discussed. In Section 4, we discuss the problem of informative prior specification for the regression parameters in the p-regression. In Section 5, we recall the Metropolis-adjusted Langevin algorithm and describe its application in the current context. In Section 6, we present the analysis of our burrow count dataset.

2. Zero-inflated modeling

Given a parametric distribution $\pi(y|\theta)$ on the integers $y = 0, 1, 2, \ldots,$ we define the associated zero inflated distribution to have

$$P(Y = 0|p, \theta) = p + (1 - p)\pi(0|\theta)$$
$$P(Y = y|p, \theta) = (1 - p)\pi(y|\theta), y > 0. \tag{1}$$

That is, a proportion p is taken from $\pi(y|\theta)$ and assigned to the event $\{y = 0\}$; evidently $P(Y = 0|p, \theta) > \pi(0|\theta)$. Note that the idea in (1) can be used to inflate the probability at any integer but, in practice, 0 is usually the value of interest. Candidates for $\pi(y|\theta)$ include the Poisson, negative binomial, binomial, beta binomial and hypergeometric distributions. For future use we note that $P(Y = y > 0|Y > 0, p, \theta) = \pi(y|\theta)/(1 - \pi(0|\theta))$, free of p. Also, $E(Y|p, \theta) = (1 - p)E_\pi(Y|\theta)$, and $\mathrm{var}\,(Y|p, \theta) = p(1 - p)(E_\pi(Y|\theta))^2 + (1 - p)\,\mathrm{var}_\pi(Y|\theta)$ where $E_\pi(Y|\theta)$ and $\mathrm{var}_\pi(Y|\theta)$ denotes the expectation and variance of a random variable Y with probability mass function π.

Introduction of a latent indicator variable facilitates working with (1). Define the joint distribution for Y and Z as follows: $P(Y = 0, Z = 1|p, \theta) = p, P(Y = y, Z = 0|p, \theta) = (1 - p)\pi(y|\theta)$. Then, marginally Z is a Bernoulli random variable with success probability p. Conditionally, $P(Y = 0|Z = 1) = 1, P(Y = y|Z = 0, p, \theta) = \pi(y|\theta), P(Z = 1|Y = y > 0) = 0, P(Z = 1|Y = 0, p, \theta) = p/(p + (1 - p)\pi(0|\theta))$.

For a sample of size n with Y_i given p_i and θ_i distributed as in (1), the full data likelihood arises from $\prod_i P(Y_i = y_i|Z_i = z_i)P(Z_i = z_i)$ and takes the form

$$L(\boldsymbol{p}, \boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{Z}) = \prod_{i=1}^{n} p_i^{z_i}((1 - p_i)\pi(y_i \mid \theta_i))^{1 - z_i}. \tag{2}$$

In fact, because Z_i is degenerate at 0 if $Y_i > 0$, the only nondegenerate Z_i (and thus the only ones we need to introduce) are those associated with the $Y_i = 0$. Therefore, (2) can be rewritten as

$$\prod_{y_i > 0}(1 - p_i)\pi(y_i|\theta_i) \prod_{y_i = 0} p_i^{z_i}((1 - p_i)\pi(0|\theta_i))^{1 - z_i}. \tag{3}$$

The marginal or observed data likelihood takes the form

$$L(\boldsymbol{p}, \boldsymbol{\theta}; \boldsymbol{Y}) = \prod_{i=1}^{n}\{p_i 1(y_i = 0) + (1 - p_i)\pi(y_i|\theta_i)\}. \tag{4}$$

In the i.i.d. case $\log L(p, \theta; \boldsymbol{Y}) = V_0 \log(p(1 - \pi(0|\theta)) + \pi(0|\theta)) + (n - V_0)\log(1 - p) + \sum_i \log \pi(y_i|\theta)$, where $V_0 = \#(Y_i = 0)$, which is immediately unimodal in p for each θ. In the literature, the most frequent choice for $\pi(y|\theta)$ is a Poisson(λ) leading to the ZIP(p, λ) models mentioned in the introduction. Here $E(Y|p, \lambda) = (1 - p)\lambda < \lambda$ and $\mathrm{var}(Y|p, \lambda) = (1 - p)\lambda(1 + p\lambda) > E(Y|p, \lambda)$. The ZIP($p, \lambda$) models are overdispersed relative to the Poisson(λ) model. The negative binomial model (without zero inflation), $P(Y = y|p, r) = (y + r - 1/r - 1)p^r(1 - p)^y, y = 0, 1, 2, \ldots,$ offers a comparably rich alternative two-parameter family to the ZIP(p, λ). However, modeling the integer valued parameter r using covariates is awkward; modeling the mean through covariates is difficult due to the combinatoric form. By contrast, the ZIP(p, λ) is easily fitted using the latent Z_i

and the parameters p and λ are familiar to model. Hence we turn to the ZIP regression set up.

Let $Y_i \sim \text{ZIP}(p_i, \lambda_i)$ where Y_i's are independently distributed. Using canonical links, we assume the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\boldsymbol{p} = (p_1, \ldots, p_n)'$ satisfy

$$\log(\boldsymbol{\lambda}) = \boldsymbol{B\beta},\ \text{logit}(\boldsymbol{p}) = \log\left(\frac{\boldsymbol{p}}{\boldsymbol{1}-\boldsymbol{p}}\right) = \boldsymbol{G\alpha},$$

where $\quad \boldsymbol{B} = \begin{pmatrix} b_1' \\ \cdot \\ \cdot \\ b_n' \end{pmatrix} \quad$ and $\quad \boldsymbol{G} = \begin{pmatrix} g_1' \\ \cdot \\ \cdot \\ g_n' \end{pmatrix}$

are specified design matrices with $\beta = (\beta_0, \beta_1, \ldots, \beta_{(q-1)})'$ and $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{(m-1)})'$ the associated parameter vectors. Following Lambert (1992), we refer to this structure as a ZIP regression. B and G may share some common covariates but need not be the same. Lambert details the use of the E-M algorithm working with (2) in order to provide likelihood-based inference for (4). If we assume that $\boldsymbol{B} = \boldsymbol{G}$ and $\boldsymbol{p}$ is a function of λ, we can reduce the number of parameters in the model. A natural parametrization in this case is $\log(\lambda) = \boldsymbol{B}\beta$ and logit $(\boldsymbol{p}) = -\kappa \boldsymbol{B}\beta$ where κ is an unknown real-valued shape parameter. Lambert notes that in this case likelihood-based inference is more easily implemented using a Newton-Raphson algorithm. Ghosh *et al.* (1998) discuss inference using a Bayesian formulation in the case where $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are unrelated. To capture additional heterogeneity in the Y_is we can assume (4) takes the form

$$\log(\boldsymbol{\lambda}) = \boldsymbol{B\beta} + \boldsymbol{W}_{1\varphi}, \text{logit}(\boldsymbol{p}) = \log\left(\frac{\boldsymbol{p}}{\boldsymbol{1}-\boldsymbol{p}}\right) = \boldsymbol{G\alpha} + \boldsymbol{W}_{2\gamma}, \tag{5}$$

where $\boldsymbol{\varphi}$, $\boldsymbol{\gamma}$ are random effects and $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ are appropriate incidence matrices. If $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are not related, it is reasonable to assume that $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ are independent of each other. Conditionally on $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ (as well as $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$) the Y_is are independent. Marginalizing over $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ they are not. So (5) provides a specification for modeling dependencies in zero-inflated counts. Inference using sampling based methods (in particular Gibbs sampling) under a fully Bayesian formulation provides an attractive approach for fitting (5). Samples obtained from the posterior distribution of ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\gamma}$) provide exact inference for the parameters (or functions of the parameters) without relying on numerical or asymptotic approximations.

As noted in the introduction, spatial count data arise frequently in practice. Spatial association between the counts is most easily modeled using (5) by introducing spatial random effects. With lattice or areal unit data a Markov random field specification using a conditionally autoregressive (CAR) model (Besag, 1974; Cressie, 1993, pp. 430–441) is commonly employed. With geostatistical or point referenced data, these random effects would be assumed to arise from a suitable spatial process (as in Diggle *et al.*, 1999). Since our experimental data arise on a regular grid of pixels, we focus on the CAR case in the sequel. Regardless, there is the question of where to introduce the spatial effects in (5). Recalling the joint distribution for Y_i and Z_i above, the model for $\log \lambda_i$ arises from $P(Y_i = y_i | Z_i = 0)$ while the model for $\log(p_i/1 - p_i)$ arises from $P(Z_i = 1)$. Hence, to explain spatial association in the observed counts, we introduce a spatial effect for each λ_i. Were most of the Z_is observed, we could introduce a spatial effect for each p_i. However, with the high proportion of 0s which motivate our ZIP modeling, most of the Z_is (82% in our example) are unobserved. Introducing a random effect, spatial or otherwise, into each p_i will provide unstable model fitting. Hence in (5) we set $\boldsymbol{W}_1 = I, \boldsymbol{W}_2 = \boldsymbol{0}$.

3. Posterior propriety

The full Bayesian model arising from (2) and (4) may be expressed in the form

$$\prod_i f_{Y|Z,\boldsymbol{\beta}}(Y_i|Z_i, \boldsymbol{\beta}) \prod_i f_{Z|\boldsymbol{\alpha}}(Z_i|\boldsymbol{\alpha}) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) f_{\boldsymbol{\beta}}(\boldsymbol{\beta}), \tag{6}$$

with resulting posterior for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{Z}$ given $\boldsymbol{Y}$ proportional to (6). Here $\boldsymbol{Z}$ denotes the vector of all unobserved Zs. Since prior information regarding $\boldsymbol{\beta}$ and especially $\boldsymbol{\alpha}$ is likely weak, improper priors are often proposed for $\boldsymbol{\alpha}$ and/or $\boldsymbol{\beta}$. Thus, the question of posterior propriety arises under (6).

We note that given $\boldsymbol{Z}$ and $\boldsymbol{Y}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent. Therefore,

$$f_{\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{Y}}(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{Y}) = \sum_z f_{\boldsymbol{\alpha}|\boldsymbol{Z}}(\boldsymbol{\alpha}|\boldsymbol{Z}) f_{\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{\beta}|\boldsymbol{Z}, \boldsymbol{Y}), \tag{7}$$

where the sum in (7) is over all possible configurations of the unobserved z. Hence propriety of (7) requires propriety of $f_{\boldsymbol{\alpha}|\boldsymbol{Z}}(\,\cdot\,|\boldsymbol{Z})$ and $f_{\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{Y}}(\,\cdot\,|\boldsymbol{Z}, \boldsymbol{Y})$ for each possible z. In the ZIP case, the former is the posterior associated with a binary logistic regression, the latter is the posterior associated with a Poisson regression.

Work dating to Diaconis and Ylvisaker (1979) has addressed the propriety of posteriors associated with generalized linear models (GLMs) under both flat and Jefferys' prior specifications. The paper of Ibrahim and Laud (1991) is noteworthy here, as is a survey paper of Gelfand and Ghosh (2000). For instance, under a flat prior for $\boldsymbol{\beta}$ which is $q \times 1$, Diaconis and Ylvisaker (1979) show that in the Poisson case a proper posterior for $\boldsymbol{\beta}$ is only assured if at least q of the Y_is are strictly positive. In the binary regression they note that a flat prior on $\boldsymbol{\alpha}$ need not give a proper posterior. (The case where $\boldsymbol{\alpha}$ is just a common intercept and all the Zs are 0 or all of the Zs are 1 makes this clear.) Under Jeffreys' prior, Ibrahim and Laud (1991) reduce the problem of posterior propriety to checking the existence of a set of one dimensional integrals. But, in general, propriety is not assured.

Hence to avoid propriety concerns we take $f_{\boldsymbol{\alpha}}(\,\cdot\,)$ and $f_{\boldsymbol{\beta}}(\,\cdot\,)$ proper in the sequel. In fact, employing simulation to fit (6), fairly informative priors are required to achieve a well-behaved Gibbs sampler.

Finally, if $\prod_i f_{Y|Z,\boldsymbol{\beta}}(Y_i|Z_i, \boldsymbol{\beta})$ is extended to include spatial random effects, modeled using a CAR prior with precision τ^2, (6) is extended to

$$\prod_i f_{Y|Z,\boldsymbol{\beta},\varphi}(Y_i|Z_i, \boldsymbol{\beta}, \varphi_i) \prod_i f_{Z|\boldsymbol{\alpha}}(Z_i|\boldsymbol{\alpha}) f_{\boldsymbol{\varphi}|\tau^2}(\boldsymbol{\varphi}|\tau^2) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) f_{\boldsymbol{\beta}}(\boldsymbol{\beta}) f_{\tau^2}(\tau^2). \tag{8}$$

The usual CAR prior $f_{\boldsymbol{\varphi}|\tau^2}(\boldsymbol{\varphi}|\tau^2)$ is improper. Following Besag *et al.* (1995) as a remedy we impose the constraint $\sum_i \varphi_i = 0$ after each iteration of the Gibbs sampler.

4. Proper prior specifications

Following the discussion of the previous section, in (6) we employ proper priors for $\boldsymbol{\beta}$ and for $\boldsymbol{\alpha}$. In fact, we suggest somewhat noninformative multivariate normal specifications in each case. By somewhat noninformative we propose setting the mean equal to $\mathbf{0}$ and identifying an appropriate order of magnitude for the variability.

For $\boldsymbol{\beta}$ this is straightforward. Fit a Poisson regression with a standard software package using only the $Y_i's > 0$ to obtain say $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{(1)}$, the asymptotic covariance matrix associated with the estimated $\boldsymbol{\beta}$. Do the same using all of the Y_is to obtain say $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{(2)}$. Let $D_{\boldsymbol{\beta}}$ be the diagonal matrix whose entries are the diagonal elements of $(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{(1)} + \hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{(2)})/2$. Adopt as the prior for $\boldsymbol{\beta}, \mathrm{N}(\mathbf{0}, kD_{\boldsymbol{\beta}}), \boldsymbol{k} > 1$ (eventually experimenting with several choices of $\boldsymbol{k}$ to assess prior sensitivity).

For $\boldsymbol{\alpha}$, with the Z_is unobserved, there is no binary regression analog of the above available. Suppose, we could obtain a simulated set of Z_is. Then we could use standard software to obtain $\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}}^{()}$, the asymptotic covariance matrix associated with the estimated $\boldsymbol{\alpha}$ and, letting $D_{\boldsymbol{\alpha}}$ be the diagonal matrix associated with $\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}}^{()}$, we could adopt as the prior for $\boldsymbol{\alpha}$, $N(\mathbf{0}, kD_{\boldsymbol{\alpha}})$ (again experimenting with k). In fact, we would simulate several sets of Z_is to obtain a better feeling for the order of the elements of $D_{\boldsymbol{\alpha}}$.

How shall we simulate the Z_is? The Z_is will be most variable if $\mathrm{P}(Z_i = 1) = 0.5$ for each i. However, maximizing variability in the Z_is ignoring $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ does not imply large variability in the $\hat{\alpha}$s when a logistic regression is fitted to these Z_is. Instead, if the p_is are assumed constant, the Z_is are conditionally independent with $P(Z_i = 1|Y_i = 0, p, \boldsymbol{\beta}) = p/(p + (1-p)\exp(-\exp(b_i^T\boldsymbol{\beta})))$. Using $\hat{\beta}$ from the Poisson regression, with an estimate of p we could simulate each of the Z_is.

Cheap estimates of p can be obtained in several ways. For example, if we set all $\lambda_i = \lambda$ then, recalling that V_0 denoted $\#(Y_i = 0)$, V_0/n estimates $p + (1-p)\mathrm{e}^{-\lambda}$. But also $\bar{Y} = \sum Y_i/(n - V_0)$, the average of the positive Y_is, estimates $E(Y|Y > 0, \lambda) = (1 - \mathrm{e}^{-\lambda})^{-1}\lambda$. These two equations provide a ''method of moments'' estimate of p. Alternatively, under (2) and (5) $P(Y_i = 2|\boldsymbol{\alpha}, \boldsymbol{\beta})/P(Y_i = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp(b_i^T\boldsymbol{\beta})/2$. But also $P(Y_i = 1|\boldsymbol{\alpha}, \boldsymbol{\beta})/(P(Y_i = 0|\boldsymbol{\alpha}, \boldsymbol{\beta}) - p) = \exp(b_i^T\boldsymbol{\beta})$. So, regardless of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $p = P(Y_i = 0|\boldsymbol{\alpha}, \boldsymbol{\beta}) - (P(Y_i = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}))^2/(2P(Y_i = 2|\boldsymbol{\alpha}, \boldsymbol{\beta}))$. As a first order approximation, $\hat{p} = V_0/n - V_1^2/2nV_2$ where $V_l = (Y_i = l)$. In our limited experience, each of the above methods has given the same order of variability for $\boldsymbol{\alpha}$.

Lastly, the code of Lambert (other versions are available, e.g., from Peter Perkins, peter@caliban.ucsd.edu) will provide $\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}}^{()}$ arising through the EM algorithm which can be used to obtain $D_{\boldsymbol{\alpha}}$. The resultant variability in $\hat{\boldsymbol{\alpha}}$ is smaller than that of the foregoing methods, roughly by an order of 10. So, to weaken prior assumptions, we employed the order of the previous paragraph. In fact, we further inflated prior variability for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by setting $k = 10$ above.

5. Implementation issues

In fitting (8) we must update $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\varphi}$, $\boldsymbol{Z}$ and τ^2. With an inverse Gamma prior on τ^2, the resulting full conditional distribution will also be inverse Gamma. The full conditionals for the Z_is are updated Bernoullis. For the φ_i the CAR specification along with the likelihood conveniently gives the form for the full conditionals. The full conditional distributions of the αs are log concave and hence sampled using the adaptive rejection sampling procedure (Gilks and Wild, 1992). In updating $\boldsymbol{\beta}$ we have had good success with Metropolis steps using a Langevin diffusion approach to provide a proposal density.

In particular, suppose π is an arbitrary density function. Assume π is everywhere non-zero and differentiable so that $\bigtriangledown \log(\pi(x))$ is well defined. Then the Langevin diffusion L_t is defined by the m-dimensional stochastic differential equation $dL_t = dW_t + 1/2 \bigtriangledown \log \pi(L_t)dt$ where $\boldsymbol{W}_t$ is m-dimensional standard Brownian motion. When π is sufficiently smooth, it can be shown that L_t has π as a stationary measure, and also that $||P_L^t(x; A) - \pi(A)|| \to 0$, a.e. x where $P_L^t(x; A) = P(L_t \in A | L_0 = x)$. Roberts and Tweedie (1996) study conditions under which such convergence is exponentially fast. For implementation purposes, one has to find good discrete approximations to L_t converging to π. We employed a Metropolis-adjusted chain denoted generically by M_n which proposes to update U_n according to $U_n | M_{n-1} \sim N(M_{n-1} + 1/2h \bigtriangledown \log \pi(M_{n-1}), hI_m)$. U_n is accepted with probability $\alpha(M_{n-1}, U_n) = \min(1, \pi(U_n)q(U_n, M_{n-1})/\pi(M_{n-1})q(M_{n-1}, U_n))$, where $q(U_n, M_{n-1})$ is the above normal density evaluated at U_n given M_{n-1}. This Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996) can be used to sample full conditionals within the Gibbs sampler provided they are differentiable and non-vanishing. Generally, the algorithm has high optimal acceptance rates even for high dimensions; the optimal acceptance rate converges to 57.8% as the dimension goes to infinity, compared with the 22% acceptance rate for simple random walk algorithms (Roberts and Rosenthal, 1995).

Specifically, the full conditional density for $\boldsymbol{\beta}$ is proportional to $\exp(-\sum_i (1 - Z_i) \exp(\varphi_i + b_i^T \boldsymbol{\beta}) + \sum_i Y_i b_i^T \boldsymbol{\beta})$. To choose the discretization step h, we ran a Poisson regression using only the positive counts. The variance of $\boldsymbol{\beta}$ gave us an idea regarding the scale of h. We then adjusted the value of h until we achieved an acceptance rate close to 70%. The initial values for the β's in the Gibbs sampler were determined by the Poisson regression. The initial values for the α's were determined by logistic regression as discussed in Section 4. We reran the Gibbs sampler using overdispersed initial values. The point estimates and credible intervals did not change significantly. We obtained a well behaved Markov chain with little autocorrelation. A burn in sample of 2000 was adequate in all cases.

6. Data analysis

With regard to the dataset mentioned in the introduction, the broad goal is to enhance our understanding of why a particular animal chooses to live in one location rather than another. The particular organism used in this study is a terrestrial isopod, *Hemilepistus reaumuri*. Its life history is well studied. See, for example, Baker *et al.* (1998), Citron-Pousty and Shachak (1998) and further references therein. These isopods live in family burrows which provide shelter and humidity in the arid environment. Previous studies have shown that the isopods respond to rainfall runoff redistribution (Yair and Shachak, 1982; Shachak and Yair, 1984; Shachak and Brand, 1991). Areas with increased runoff tend to have increased probability of settling (Citron-Pousty and Shachak, 1998). Because of the dry conditions, water is a limiting factor in the environment.

Small first order watersheds were surveyed in the Haluqim ridge of the Negev desert, Israel. In this desert, dewfall can contribute relatively large amounts of water per year and has less annual variability when compared to rainfall (Zangvil, 1996; Kidron, 1998). Our watershed opened in a westerly direction. The watershed was exhaustively surveyed for burrows in the summer of 1995. All burrow locations were mapped in the Israel coordinate system.

For this watershed we chose a pixel size of $5\,\mathrm{m} \times 5\,\mathrm{m}$ and created four raster GIS data layers: (i) counts per pixel; (ii) average time of dew availability in the pixel; (iii) percentage shrub cover in the pixel; and (iv) percentage rock cover in the pixel. There was only one other cover type, percentage soil, so the latter layers uniquely determine the coverage percentages for a pixel. Based upon previous research we expect a positive relationship with regard to burrow survival (hence desirability of the site) for each of dew availability, percentage rock, and percentage shrub.

Fig. 1 shows the region under study with the burrow counts overlaid. We have a total of 1649 pixels of which 82.1% have zero counts. The explanatory variable dew duration measures time (in hundredths of an hour from 8 a.m.) to evaporation of dew. Fig. 2 shows the distribution of dew duration over the region. The explanatory variables shrub density and rock density measure the abundance of shrub and rock respectively at a given pixel. At a resolution of $0.1\,\mathrm{m} \times 0.1\,\mathrm{m}$, we have binary maps showing shrub and rock incidences. These were aggregated to a resolution of $5\,\mathrm{m} \times 5\,\mathrm{m}$ giving us the shrub and rock densities (as a percent). Figs 3 and 4 show the distribution of shrub and rock density over the region.

We fitted the spatial ZIP regression model in (8) after centering (at 0) and scaling (using the sample standard deviation) all the covariates. The prior specification for $\boldsymbol{\beta}$ (which includes the intercept and then the coefficients) is $N(\mathbf{0}, 10D_{\boldsymbol{\beta}})$ where $(D_{\boldsymbol{\beta}})_{11} = 0.004$, $(D_{\boldsymbol{\beta}})_{22} = 0.001$, $(D_{\boldsymbol{\beta}})_{33} = 0.005$ and $(D_{\boldsymbol{\beta}})_{44} = 0.007$. We obtained priors for $\boldsymbol{\alpha}$ using all the methods discussed in Section 4 and finally used the one that provided the largest variability; $\boldsymbol{\alpha} \sim N(\mathbf{0},\ 10D_{\boldsymbol{\alpha}})$ where $(D_{\boldsymbol{\alpha}})_{11} = 0.004$, $(D_{\boldsymbol{\alpha}})_{22} = 0.004$, $(D_{\boldsymbol{\alpha}})_{33} = 0.012$ and $(D_{\boldsymbol{\alpha}})_{44} = 0.013$. For the spatial variability parameter τ^2, we adopt an inverse Gamma prior. In particular $\tau^2 \sim \mathrm{IG}\ (2, 0.8)$. This distribution has infinite variance and mean roughly the sample variability in the $\log Y_i$ (where $Y_i > 0$). Table 1 provides the posterior median and 95% credible intervals for the regression parameters and the spatial variability parameter τ^2.

Table 1. Parameter estimation (point and interval estimates) for spatial ZIP regression model without trend surface.

Quantiles	*2.5%*	*50%*	*97.5%*
Parameters for Poisson regression			
β_1 (dew duration)	0.077	0.209	0.333
β_2 (shrub density)	0.539	0.724	0.897
β_3 (rock density)	0.129	0.344	0.600
τ^2 (spatial variance)	1.97	2.58	3.33
Parameters for logistic regression			
α_1 (dew duration)	− 0.526	− 0.300	− 0.091
α_2 (shrub density)	− 1.08	− 0.804	− 0.526
α_3 (rock density)	− 1.55	− 1.18	− 0.770

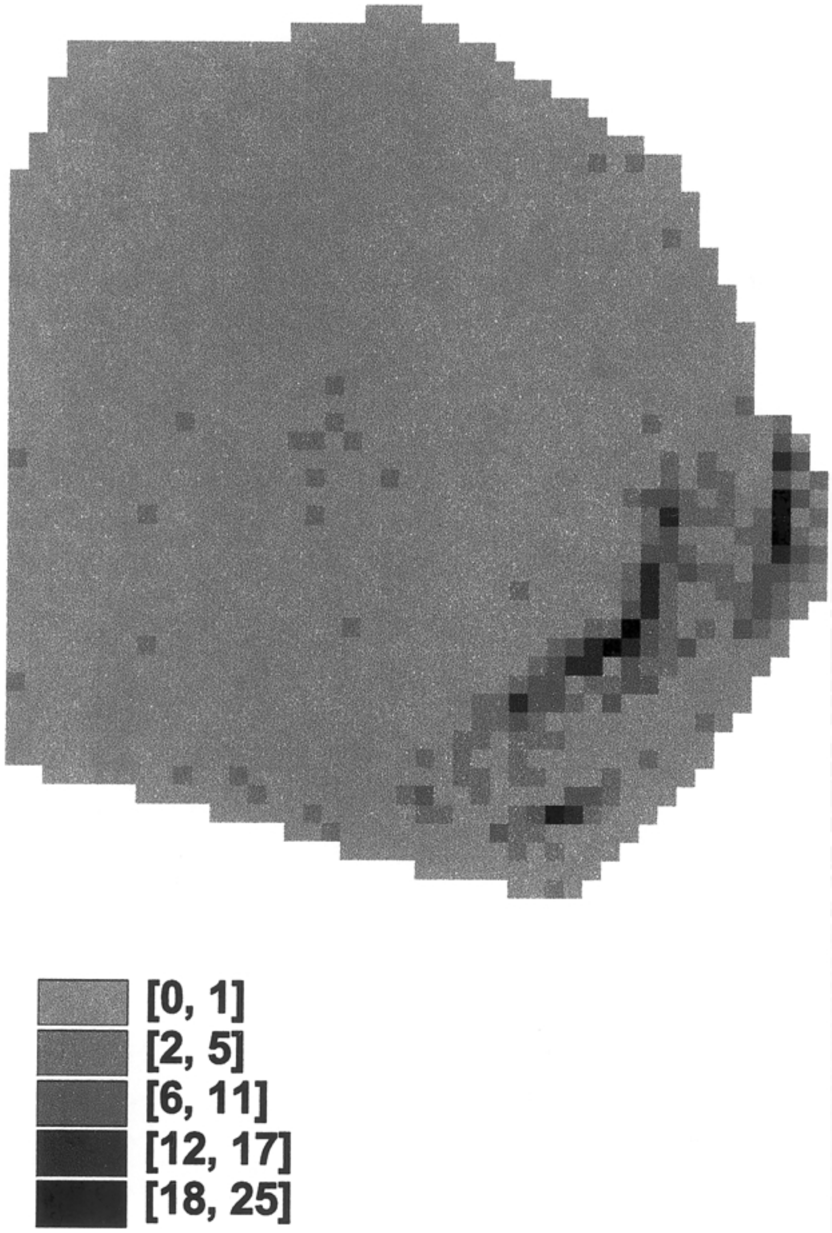

Figure 1. Burrow counts for area under study.

Dew, shrub, and rock turn out to be significant in explaining the λ_is. As expected, high dew content in the soil, presence of shrub, and presence of rock all encourage isopod settlements. As far as the p_is are concerned, again dew, shrub, and rock turn out to be significant in explaining the p_is. Isopods tend to be absent under low dew content in the soil, low shrub density, and low rock density.

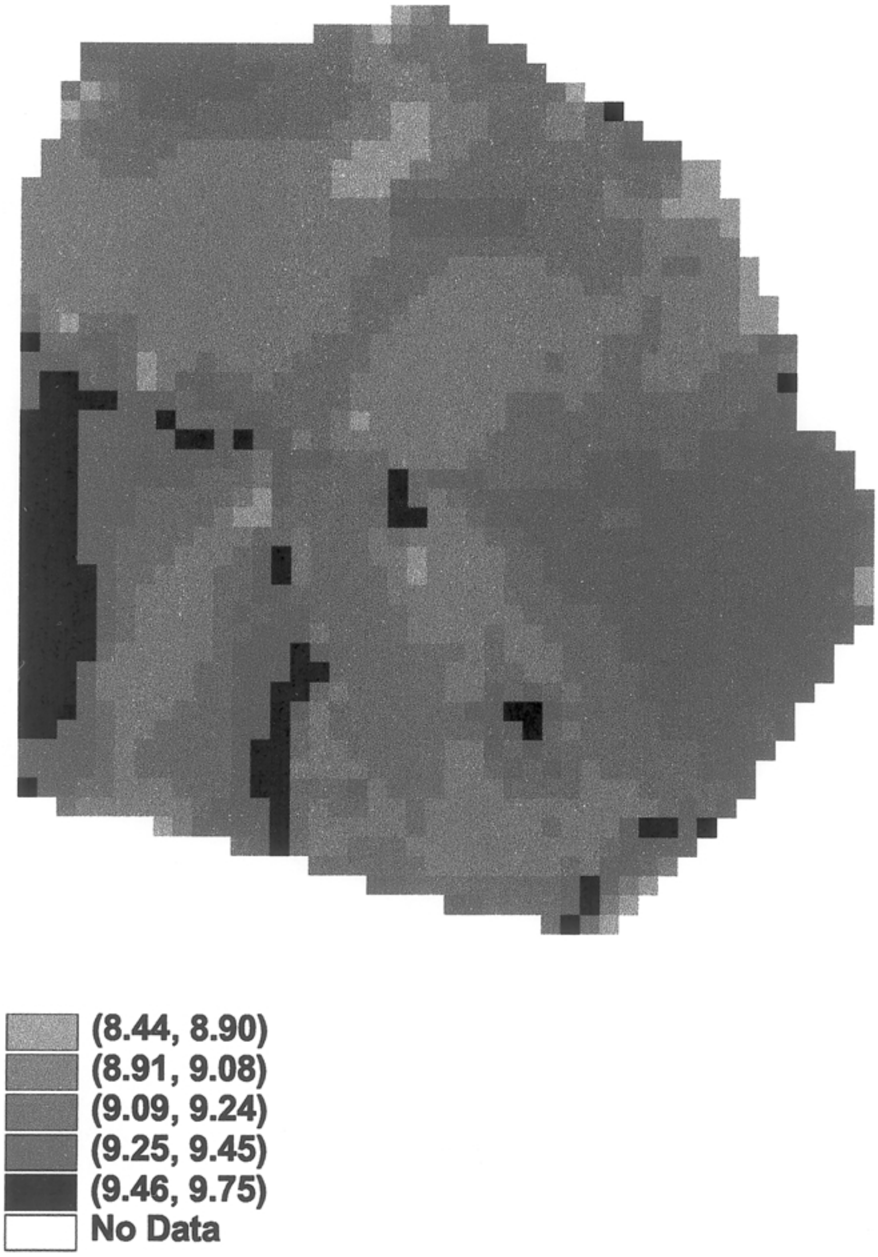

Figure 2. Distribution of dew duration for the region.

To capture the spatial picture, Fig. 5(a) presents a gray scale map of the posterior mean of the φ_is. The central interval represents effects within 0.5 standard deviation from 0. Adjacent intervals are within 0.5 to 1.5 standard deviations from 0. Extreme intervals are beyond 1.5 standard deviations. Spatial pattern is evident. Comparison with Fig. 1 shows

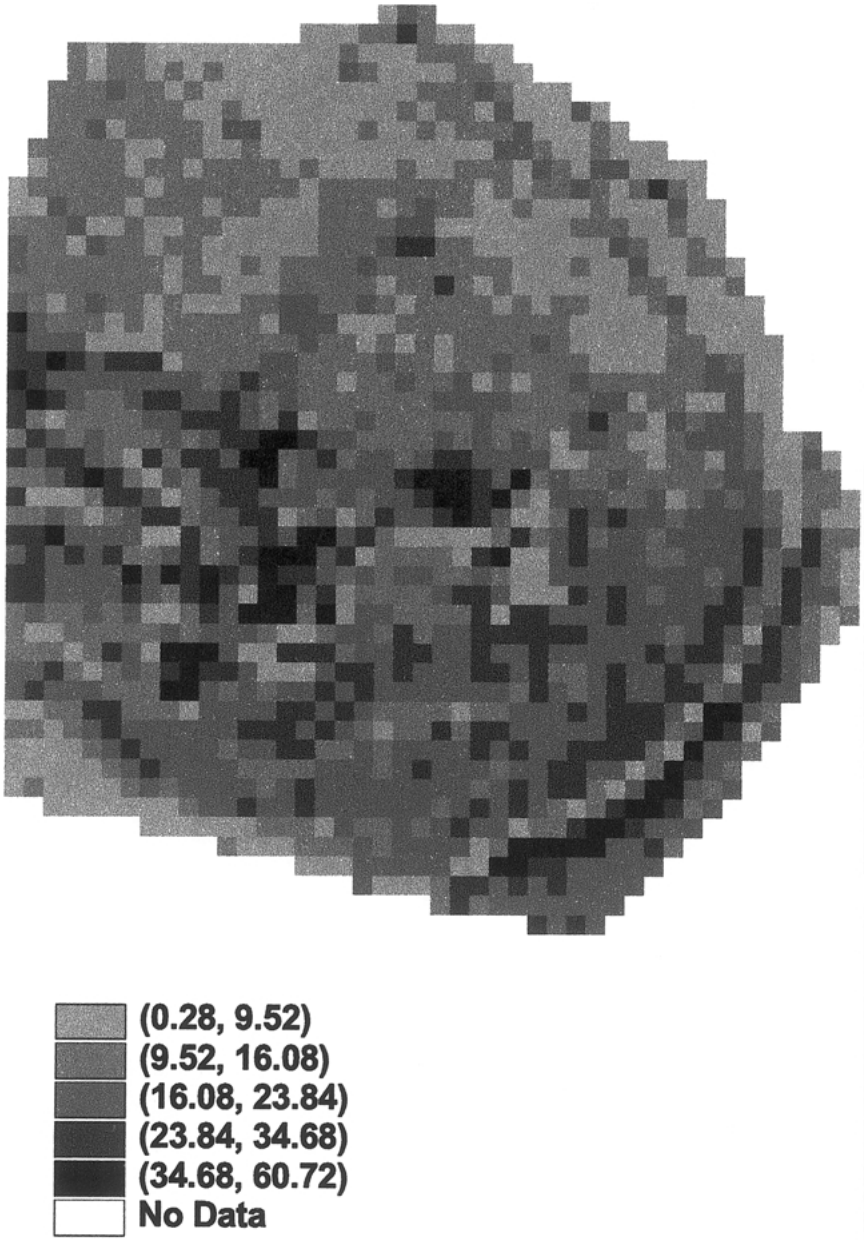

Figure 3. Distribution of shrub density (in percent) for the region.

that, using dew, shrub, and rock, underestimation is more common for pixels with higher burrow counts. Positive spatial adjustment tends to occur with larger λs.

The west, and in fact also somewhat north, opening of our watershed suggests that a trend surface introducing a coefficient for latitude and a coefficient for longitude into the Poisson regression would be of interest to explore. So, we added a β_4 and β_5 to this

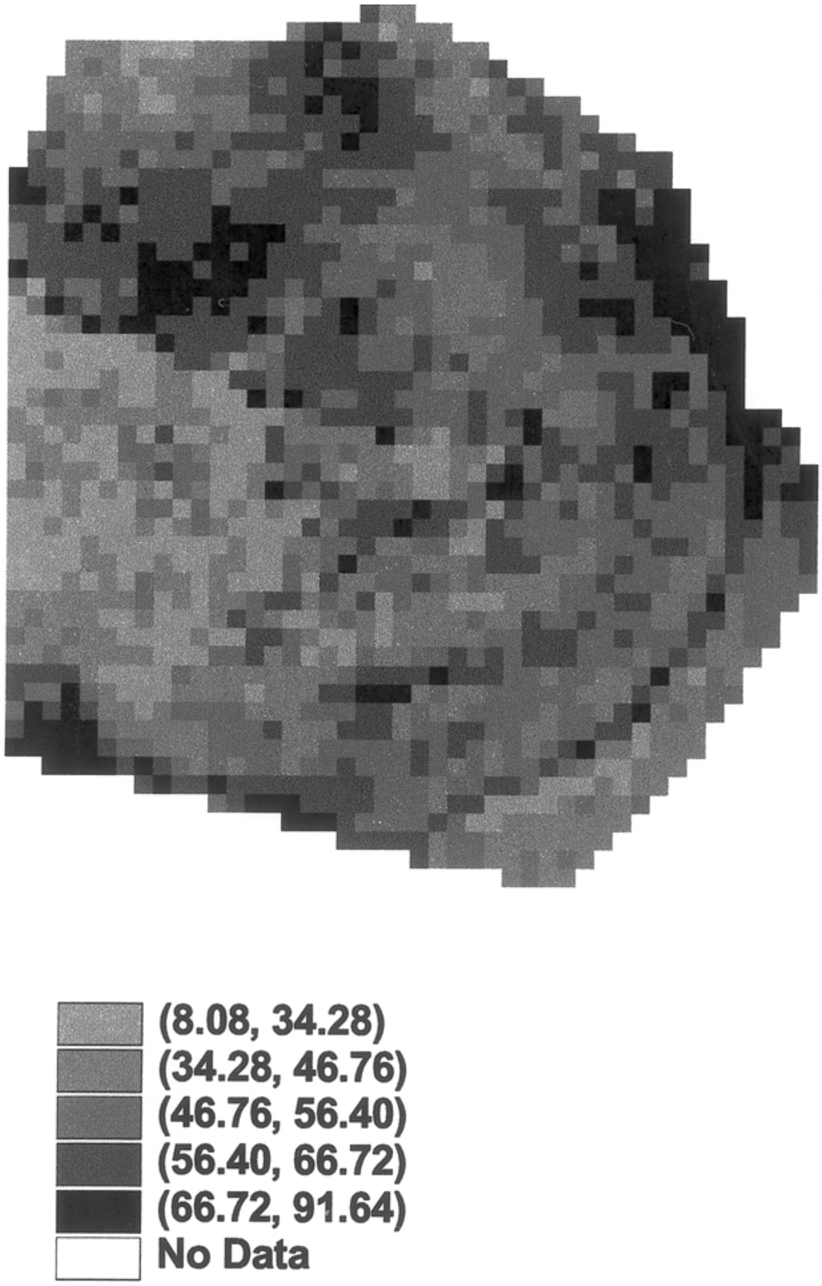

Figure 4. Distribution of rock density (in percent) for the region.

regression, centering and scaling the latitudes and longitudes. Developing priors for β_4 and β_5 as described above, $D_{\boldsymbol{\beta}}$ is extended to $(D_{\boldsymbol{\beta}})_{55} = 0.003$ (latitude) $= (D_{\boldsymbol{\beta}})_{66}$ (longitude). All previous priors are unchanged. Table 2 provides posterior summaries, analogous to those of Table 1.

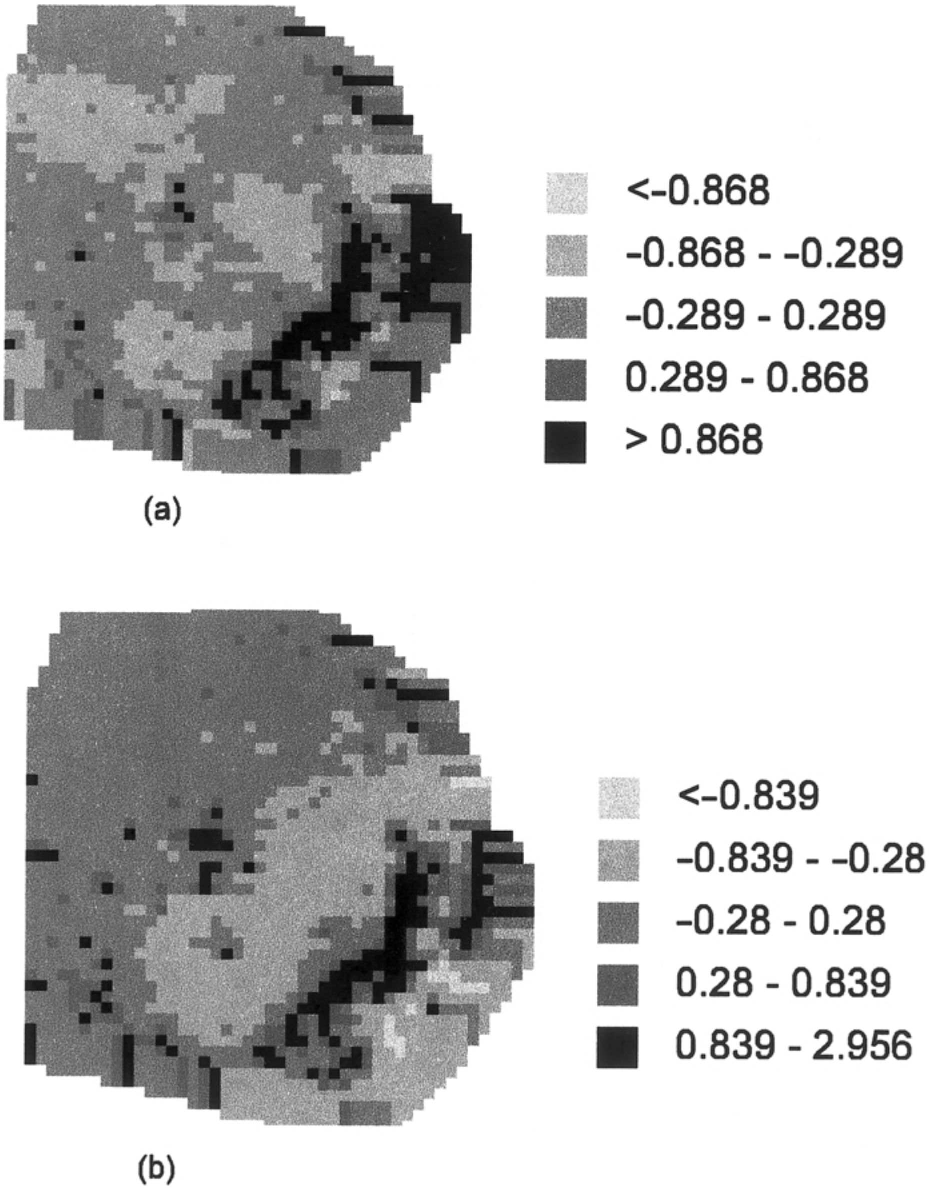

Figure 5. Posterior means for the spatial random effects. 5(a) is model without trend surface; 5(b) is model with trend surface. See text for details.

Note that $\beta_4 > 0, \beta_5 < 0$. That is, higher burrow counts are encouraged as we move east and south, in agreement with Fig. 1 and the geography of the watershed. β_1, β_2 and β_3 remain significant, roughly of the same magnitudes as before. Most interesting is that now none of the coefficients in the logistic regression for p are significant; p can be taken as constant over the region. The number of zeros in the region is still inflated relative to a Poisson model but their incidence is effectively explained by their locations. Fig. 5(b)

Table 2. Parameter estimation (point and interval estimates) for spatial ZIP regression model with trend surface.

Quantiles	*2.5%*	*50%*	*97.5%*
Parameters for Poisson regression			
β_1 (dew duration)	0.049	0.157	0.254
β_2 (shrub density)	0.481	0.677	0.838
β_3 (rock density)	0.337	0.528	0.722
β_4 (latitude)	0.709	0.829	0.961
β_5 (longitude)	− 0.738	− 0.619	− 0.482
τ^2 (spatial variance)	2.42	3.06	4.04
Parameters for logistic regression			
α_1 (dew duration)	− 0.489	− 0.216	0.085
α_2 (shrub density)	− 0.532	− 0.152	0.211
α_3 (rock density)	− 0.594	− 0.246	0.154

presents a grey scale map of the posterior means of the φ_is for this model. Spatial pattern still remains but is less detailed than in Fig. 5(a).

Acknowledgments

The work of Deepak K. Agarwal and Alan E. Gelfand was supported in part by NSF-DMS 99-71206. The work of Steven Citron-Pousty was supported by the Connecticut Space Grant Consortium, The Blaustein Institute for Desert Research, and the National Science Foundation through Grants DEB-9707785 and INT-9714044. The authors thank three referees for valuable suggestions which improved the presentation of the paper.

References

Baker, M.B., Shachak, M., Brand, S., and Yair, A. (1998) Settling behavior of the desert isopod *Hemilepistus reaumuri* in response to variation in soil moisture and other environmental cues. *Israel Journal of Zoology*, **44**, 345–54.

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

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

Citron-Pousty, S. and Shachak, M. (1998) Comparative spatial patterns of the terrestrial isopod *Hemilepistus reaumuri* at multiple scales. *Israel Journal of Zoology*, **44**, 355–68.

Clayton, D.G., and Bernardinelli, L. (1991) Bayesian methods for mapping disease risk in Small Area Studies, in *Geographical and Environmental Epidemiology*, J. Cuzick and P. Elliot (eds.), Oxford University Press, Oxford, pp. 205–20.

Cohen, A.C. (1991) *Truncated and censored samples*, Marcel Dekker, New York.

Cressie, N.A.C. (1993) *Statistics for Spatial Data* (rev. ed.), John Wiley, New York.

Dempster, A., Laird, N., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **39**, 1–38.

Diaconis, P. and Ylvisaker, D. (1979) Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269–81.
Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.
Gelfand, A.E. and Ghosh, M. (2000) *Generalized Linear Models: A Bayesian Perspective*, D.K. Dey, S.K. Ghosh and B.K. Mallick (eds.), Marcel Dekker, Inc., New York, pp. 3–14.
Gilks, W.R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–48.
Ghosh, S.K., Mukhopadhyay, P., and Lu, J.C. (1998) Bayesian analysis of zero-inflated regression models. Technical Report, Department of Statistics, North Carolina State University.
Heilbron, D.C. (1994) Zero-altered and other regression models for count data with added zeroes. *Biometrika*, **36**(5), 531–47.
Ibrahim, J.G. and Laud, P.W. (1991) On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, **86**, 981–6.
Johnson, N.L. and Kotz, S. (1969) *Distributions in Statistics: Discrete Distributions*, Houghton Mifflin, Boston, MA.
Kidron, G. (1998) Dew variability, lichen and cyanobacteria distribution along slopes at Sde Boqer-northern Negev. Jerusalem. Technical Report, New Mexico State University.
Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.
Roberts, G.O. and Rosenthal, J.S. (1995) Optimal scaling of discrete approximations to Langevin diffusions. Research Report 95-11, Statistical Laboratory, Cambridge University.
Roberts, G.O. and Tweedie, R.L. (1996) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernouilli*, **2**, 341–64.
Shachak, M. and Brand, S. (1991) Relations among spatiotemporal heterogeneity, population abundance, and variability in a desert, in *Ecological Heterogeneity*, J. Kolasa and S.T.A. Pickett (eds.), Springer-Verlag, New York, pp. 202–23.
Shachak, M. and Yair, A. (1984) Population dynamics and role of Hemilepistus reaumuri (Audouin and Savigny) in a desert ecosystem. *Symposium of Zoological Society of London*, **53**, 295–314.
Yair, A. and Shachak, M. (1982) A case study of energy, water and soil flow chains in an arid ecosystem. *Oecologia*, **54**, 389–97.
Zagvil, A. (1996) Six years of dew observation in the Negev Desert, Israel. *Journal of Arid Environments*, **32**, 361–71.

Biographical sketches

Deepak K. Agarwal is Senior Technical Staff Member at AT&T Shannon Research Labs, Florham Park, NJ 07932-0971. Alan E. Gelfand is a Professor in the Institute of Statistics and Decision Sciences at Duke University, Durham, NC 27708-0251. Steven Citron-Pousty is a graduate student in the Department of Ecology and Evolutionary Biology, both at the University of Connecticut, Storrs, CT 06269.