Integration: Mathematical Theory and Applications ISSN: 1948-5972
Volume 2, Number 3

SEQUENTIAL HYPOTHESIS TESTING WITH SPATIALLY CORRELATED COUNT DATA

Judy X. Li[1], Daniel R. Jeske[2], Jesús R. Lara[3] and Mark Hoddle[3]

ABSTRACT

It is well known that sequential hypothesis test procedures can have appreciable cost savings compared to fixed sample size test plans. The first sequential hypothesis procedure was developed by Wald for one-parameter families of distributions and later extended by Bartlett to handle the case of nuisance parameters. However, Bartlett's procedure requires independent and identically distributed observations. In ecological applications, it is common for data to exhibit spatial correlations. We illustrate the existence of spatial correlations in pest count data by analyzing the spatial structure in a data set of mite counts. The goal of this paper is to show how to incorporate the existence of spatial correlation into a sequential hypothesis testing framework so that applications such as pest management can improve the accuracy of their treat or no-treat decisions.

Keywords: Sequential Hypothesis Testing, Generalized Linear Mixed Models, Integrated Likelihood.

1. INTRODUCTION

Sequential hypothesis testing procedures are often utilized within the agriculture industry as a cost effective approach to pest density assessments (Fowler and Lynch 1987, Mulekar et. al. 1993, Binns et. al. 2000, Young and Young 1998). In these applications, Wald's (1947) Sequential Probability Ratio Test (SPRT) is the most often used approach. As discussed by Wald and Wolfowitz (1948), when compared to the most efficient fixed sample size procedures, the SPRT often requires only half as many observations to be sampled, which can amount to a significant savings in the cost of sampling.

[1] Food and Drug Administration, Rockville MD, work was performed while in the Department of Statistics at University of California - Riverside; the opinions and information in this message are those of the author and do not reflect the view and policies of the U.S. Food and Drug Administration. 2 Department of Statistics, University of California - Riverside; 3 Department of Entomology, University of California - Riverside.

In pest assessment applications, the goal of Wald's SPRT is to distinguish between two simple hypotheses about a parameter θ that reflects pest density (e.g., the mean or median number of pests per sampling unit):

$$\mathrm{H}_0 : \theta = \theta_0 \qquad \mathrm{H}_1 : \theta = \theta_1 \quad (\theta_1 > \theta_0) .$$

Here, θ_0 would represent an acceptable pest density, below which no treatment intervention (e.g., spraying with insecticide or release of natural enemies) is required, and θ_1 represents an unacceptable pest density which calls for treatment in an attempt to ward off serious crop loss. A real-life limitation to the use of Wald's SPRT procedure is that it requires there be no unknown nuisance parameters (i.e., unknown parameters in the model that are not of primary interest). Many applications of Wald's SPRT eliminate nuisance parameters by replacing them with educated guesses. However, Bartlett (1946) proved in the independent and identically distributed (IID) case that a simple modification to Wald's SPRT is sufficient to preserve the type-1 (falsely reject H_0) and type-2 (falsely accept H_1) error rates of the sequential test procedure. Shah et. al. (2009) recently advocated the use of Bartlett's sequential procedure in the context of IID pest count applications.

The application we consider in this paper concerns orchards of trees which are typically organized into blocks that can be individually assessed and treated for pests individually, as necessary. Correctly understanding the spatial distribution of the pest is crucial when deciding which blocks to treat so as to minimize treatment costs and maximize treatment benefits. Neglecting spatial dependencies could result in improper judgment of pest densities that lead to incorrect decisions about the need for treatment. Spatial analyses have been previously used in the study of a diverse group of pests of agricultural importance such as lentils (Schotzko and O'Keeffe 1989), corn-alfalfa crop rotation system (Williams et al. 1992), cotton (Gozé et al. 2003), and grapes (Ifoulis and Savopoulou-Soultani 2006, Ramírez-Dávila and Porcayo-Camargo 2008). However, in all of the above studies, spatial analyses were conducted using transformed count data in an attempt to satisfy normality assumptions. This transformation approach has limitations, particularly for sparse count data (Gotway and Stroup 1997).

Generalized Linear Mixed Models (GLMM) are statistical models which are particularly useful for modeling discrete response variables (e. g., counts) that may exhibit correlation (Breslow and Clayton 1993). As an extension to the generalized linear model, a GLMM contains both fixed effects and random effects in the link function. GLMM's have been used across multiple scientific disciplines, including ecological studies of pest populations (Candy 2000, Elston et. al. 2001, Barchia et. al. 2003, Paterson and Lello 2003, Elias et. al. 2006, Bennett et. al. 2008, Bianchi et. al. 2008, Takakura 2009).

In this paper, we combine a GLMM that has spatial structure in its random effects with use of a sequential probability ratio test to test for critical pest densities. First, we describe the spatial GLMM we have in mind, and then demonstrate the usefulness of it by analyzing the spatial structure of real count data measured for the persea mite (*O. perseae*). The persea mite is an avocado leaf feeding pest that is native to Mexico and is a serious invasive pest in California (USA), Costa Rica, Israel, and Spain (Hoddle 2005). Second, we propose a pest assessment sampling methodology that is suited for contexts where periodic assessments are

to be made to determine if pest treatments are necessary. Our proposed sampling plan consists of a first occasion fixed sample followed by sequential samples on each subsequent monitoring occasion. On each sampling occasion, H_0 versus H_1 will be tested.

2. MODEL DEVELOPMENT

2.1. Spatial GLMM

The negative binomial distribution is a common distribution in pest control studies due to its flexibility in handling over dispersed count data (i.e., variance is larger than the mean). Useful for our development of a suitable spatial GLMM is recognition of an equivalent way of arriving at the commonly used negative binomial distribution. Suppose that a pest count *Y* has a Poisson distribution with mean θ. We write the probability function of *Y* as $p(y \mid \theta) = \exp(-\theta)\theta^y / y!$, for $y \in \{0, 1, \dots\}$. Suppose θ randomly varies between sampling units, following a gamma distribution with degrees of freedom κ and rate λ. We write the density function of θ as $f(\theta) = \lambda^\kappa \theta^{\kappa-1} \exp(-\lambda\theta) / \Gamma(\kappa)$, for $\mu \geq 0$ and $\kappa > 0$, respectively.

Writing $\mu = r / \lambda$, it is easy to verify that the unconditional probability function for *Y* is

$$p(y) = \frac{\Gamma(y+\kappa)}{\Gamma(y+1)\Gamma(\kappa)} \left(\frac{\kappa}{\kappa+\mu}\right)^\kappa \left(\frac{\mu}{\kappa+\mu}\right)^y, \; y \in \{0, 1, \dots\}$$

which is the classically used negative binomial distribution with mean μ and over-dispersion parameter κ. It can be shown that that the variance of *Y* is $\mu + \mu^2 / \kappa$.

While the negative binomial distribution nicely incorporates over-dispersion, it does not address spatial correlation that may exist between the observations. Beginning with the negative binomial distribution, we use another type of conditioning approach to achieve that purpose. Let Y_{ijk} be the pest count of the *k*-th sampled leaf collected from the *j*-th cardinal direction (hereafter referred to as quadrants) of the *i*-th tree, with $i = 1, \cdots, n$, $j = 1, \cdots, 4$ and $k = 1, \dots, m$. We note that it is widely appreciated that pests sometimes choose a (e.g., based on morning versus afternoon sun exposure) specific quadrant (N, S, W, E) within a tree to nest. Let $\{\gamma_j\}_{j=1}^4$ denote the fixed quadrant effect, and $\underset{\sim}{S} = (S_1, \dots, S_n)'$ denote spatially correlated random tree effects. Our proposed spatial GLMM is defined as follows

$$\begin{aligned} &a. \quad Y_{ijk} \,\big|\, S_i \overset{ind}{\underset{\sim}{}} \text{Negative Binomial } (\mu_{ijk}, \kappa) \\ &b. \quad \log(\mu_{ijk}) = \gamma_j + S_i \\ &c. \quad \underset{\sim}{S} \sim MVN(0, \sigma^2 \underset{\sim}{\Sigma}(\rho)) \end{aligned} \tag{1}$$

where μ_{ijk} is the (conditional on S_i) mean of Y_{ijk} and $\sigma^2\underset{\sim}{\Sigma}(\rho)$ is the spatial exponential covariance structure whose (i, i') element is $\exp(-d_{i,i'}/\rho)$, where ρ is a scale parameter that dictates the strength of the spatial correlation, and $d_{i,i'}$ is the Euclidean distance between the i-th and i'-th *tree. Because of equation (1a) the observations within a tree will still be modeled by the flexible over-dispersed negative binomial distribution, but because of equations (1b) and (1c) the observations between different trees are modeled as spatially correlated.*

Let $\underset{\sim}{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)'$ denote fixed quadrant parameters and $\underset{\sim}{\Theta} = (\sigma^2, \rho, \kappa)'$ denote the variance parameters of the GLMM described by (1). The integrated likelihood function is

$$L(\underset{\sim}{\gamma}, \underset{\sim}{\Theta}) = \int\cdots\int_{R^n} \left\{ \prod_{i=1}^{n}\prod_{j=1}^{4}\prod_{k=1}^{m} \frac{\Gamma(Y_{ijk}+\kappa)}{\Gamma(Y_{ijk}+1)\Gamma(\kappa)} \left[\frac{\exp(\gamma_j + s_i)}{\kappa + \exp(\gamma_j + s_i)}\right]^{Y_{ijk}} \left[\frac{\kappa}{\kappa + \exp(\gamma_j + s_i)}\right]^{\kappa} \right\} \times \varphi\left(\underset{\sim}{s}; \sigma^2\underset{\sim}{\Sigma}(\rho)\right) d\underset{\sim}{s} \qquad (2)$$

where the expression inside the braces is the product of negative binomial probabilities based on the conditional mean structure defined by equation (1). The function $\varphi\left(\underset{\sim}{s}; \sigma^2\underset{\sim}{\Sigma}(\rho)\right)$ is the n-dimensional multivariate normal density function corresponding to the random tree effects $\underset{\sim}{S}$. Because n is relatively large, evaluating the likelihood function using numerical integration techniques such as quadrature rules are not feasible. Quadrature approximations are usually only feasible in this context when the random effects are uncorrelated (i.e., no spatial correlation) in which case the integral in equation (2) would reduce to a product of one-dimensional integrals.

2.2. MODEL FITTING

The approach we take to fitting the spatial GLMM is based on the method of pseudo-likelihood that was proposed in Wolfinger and O'Connell (1993). A sketch of this method, as it applies to (1) is now given. Let $\underset{\sim}{Y}_i = (Y_{i11}, \ldots, Y_{i1m}, \ldots, Y_{i41}, \ldots, Y_{i4m})'$ denote the vector of pest counts collected from the i-th tree and denote the corresponding vector of conditional means as $\underset{\sim}{\mu}_i = (\mu_{i11}, \ldots, \mu_{i1m}, \ldots, \mu_{i41}, \ldots, \mu_{i4m})'$. Let $\underset{\sim}{Y} = (\underset{\sim}{Y}_1', \underset{\sim}{Y}_2', \ldots, \underset{\sim}{Y}_n')'$ and $\underset{\sim}{\mu} = (\underset{\sim}{\mu}_1', \underset{\sim}{\mu}_2', \ldots, \underset{\sim}{\mu}_n')'$. Denote a vector of a ones by J_a and let I_a denote the $a \times a$ identity matrix. With the usual definition of Kronecker product between two matrices, namely $A \otimes B = \left[a_{ij}B\right]$, define the $4mn \times 4$ matrix $\underset{\sim}{X} = J_n \otimes (I_4 \otimes J_m)$ and the $4mn \times 4$ matrix $\underset{\sim}{Z} = I_n \otimes J_{4m}$. Then equation (1b) can be written in vector notation as $\log \underset{\sim}{\mu} = \underset{\sim}{X}\underset{\sim}{\gamma} + \underset{\sim}{Z}\underset{\sim}{S}$.

To proceed with the pseudo-likelihood approach, we informally write $\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\mu}$, or $\underset{\sim}{e} = \underset{\sim}{Y} - \exp(\underset{\sim}{X}\underset{\sim}{\gamma} + \underset{\sim}{Z}\underset{\sim}{S})$, where the second term is understood to be a vector whose t-th element is $\exp(\underset{\sim}{x}_t'\underset{\sim}{\gamma} + \underset{\sim}{z}_t'\underset{\sim}{S})$, and where $\underset{\sim}{x}_t' = (x_{t1}, x_{t2}, x_{t3}, x_{t4})$ and $\underset{\sim}{z}_t' = (z_{t1}, z_{t2}, \ldots, z_{tn})$ denote the t-th rows of $\underset{\sim}{X}$ and $\underset{\sim}{Z}$, respectively, ($t = 1, \ldots, 4mn$). Consider a first-order Taylor expansion of each element in $\underset{\sim}{e}$ about an initial guess $(\underset{\sim}{\gamma}_0, \underset{\sim}{S}_0)$, where $\underset{\sim}{\gamma}_0 = (\gamma_{10}, \gamma_{20}, \gamma_{30}, \gamma_{40})'$ and $\underset{\sim}{S}_0 = (S_{10}, S_{20}, \ldots, S_{n0})'$. Denoting this by $\underset{\sim}{e}$, we have (in vector notation) $\hat{\underset{\sim}{e}} = \underset{\sim}{Y} - \underset{\sim}{\mu}_0 - \underset{\sim}{D}(\underset{\sim}{\mu}_0)\left(\underset{\sim}{X}\underset{\sim}{\gamma} - \underset{\sim}{X}\underset{\sim}{\gamma}_0 + \underset{\sim}{Z}\underset{\sim}{S} - \underset{\sim}{Z}\underset{\sim}{S}_0\right)$, where $\underset{\sim}{D}(\underset{\sim}{\mu}_0) = Diag(\underset{\sim}{\mu}_0)$ and $\underset{\sim}{\mu}_0 = \exp(\underset{\sim}{X}\underset{\sim}{\gamma}_0 + \underset{\sim}{Z}\underset{\sim}{S}_0)$.

A consequence of equation (1a) is that $Var(\hat{\underset{\sim}{e}} \mid \underset{\sim}{S}) = \underset{\sim}{D}(\underset{\sim}{\mu}) + \underset{\sim}{D}^2(\underset{\sim}{\mu})/\kappa$, and if $(\underset{\sim}{\gamma}_0, \underset{\sim}{S}_0)$ is close to the true value, then we will have $E(\hat{\underset{\sim}{e}} \mid \underset{\sim}{S}) \approx \underset{\sim}{0}$. The conditional distribution of $\hat{\underset{\sim}{e}}$, given $\underset{\sim}{S}$, is unknown so the pseudo-likelihood approach heuristically assumes it is approximately multivariate normal. That is, it is assumed that $\hat{\underset{\sim}{e}} \mid \underset{\sim}{S} \approx MVN\left(\underset{\sim}{0}, \underset{\sim}{D}(\underset{\sim}{\mu}) + \kappa^{-1}\underset{\sim}{D}^2(\underset{\sim}{\mu})\right)$, where the symbol $\approx$ denotes 'is approximately distributed as.' Replacing $\underset{\sim}{\mu}$ by $\underset{\sim}{\mu}_0$ in this approximation then leads to

$$\underset{\sim}{D}^{-1}(\underset{\sim}{\mu}_0)(\underset{\sim}{Y} - \underset{\sim}{\mu}_0) + \underset{\sim}{X}\underset{\sim}{\gamma}_0 + \underset{\sim}{Z}\underset{\sim}{S}_0 \mid \underset{\sim}{S} \approx MVN\left(\underset{\sim}{X}\underset{\sim}{\gamma} + \underset{\sim}{Z}\underset{\sim}{S}, \underset{\sim}{D}^{-1}(\underset{\sim}{\mu}_0) + \kappa^{-1}\underset{\sim}{I}\right).$$

Defining the so-called pseudo-response variable $\underset{\sim}{v} = \underset{\sim}{D}^{-1}(\underset{\sim}{\mu}_0)(\underset{\sim}{Y} - \underset{\sim}{\mu}_0) + \underset{\sim}{X}\underset{\sim}{\gamma}_0 + \underset{\sim}{Z}\underset{\sim}{S}_0$, we see that it can be characterized as (approximately) following a classical linear mixed model $\underset{\sim}{v} = \underset{\sim}{X}\underset{\sim}{\gamma} + \underset{\sim}{Z}\underset{\sim}{S} + \underset{\sim}{\varepsilon}$, with $\underset{\sim}{S} \sim MVN(0, \sigma^2\underset{\sim}{\Sigma}(\rho))$ and independent of $\underset{\sim}{\varepsilon} \sim MVN(\underset{\sim}{0}, \underset{\sim}{D}^{-1}(\underset{\sim}{\mu}_0) + \kappa^{-1}\underset{\sim}{I})$.

A classical linear mixed model analysis can now be used to get the next iteration $(\underset{\sim}{\gamma}_1, \underset{\sim}{S}_1)$. In particular, for a fixed $\underset{\sim}{\Theta}$, the maximum likelihood estimate of $\underset{\sim}{\gamma}$ is given by $\dot{\underset{\sim}{\gamma}}(\underset{\sim}{\Theta}) = (\underset{\sim}{X}'\underset{\sim}{H}^{-1}(\underset{\sim}{\Theta})\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{H}^{-1}(\underset{\sim}{\Theta})\underset{\sim}{v}$, where $\underset{\sim}{H}(\underset{\sim}{\Theta}) = \sigma^2\underset{\sim}{Z}\underset{\sim}{\Sigma}(\rho)\underset{\sim}{Z}' + \underset{\sim}{D}^{-1}(\underset{\sim}{\mu}_0) + \kappa^{-1}\underset{\sim}{I}$, and the best linear unbiased predictor for $\underset{\sim}{S}$ is $\dot{\underset{\sim}{S}}(\underset{\sim}{\Theta}) = \sigma^2\underset{\sim}{\Sigma}(\rho)\underset{\sim}{Z}'\underset{\sim}{H}^{-1}(\underset{\sim}{\Theta})(\underset{\sim}{v} - \underset{\sim}{X}\dot{\underset{\sim}{\gamma}}(\underset{\sim}{\Theta}))$. Then next iteration is thus $(\underset{\sim}{\gamma}_1, \underset{\sim}{S}_1) = (\dot{\underset{\sim}{\gamma}}(\hat{\underset{\sim}{\Theta}}), \dot{\underset{\sim}{S}}(\hat{\underset{\sim}{\Theta}}))$, where $\hat{\underset{\sim}{\Theta}}$ is the maximum likelihood estimate obtained by maximizing (e.g., via Newton's method) the corresponding log profile likelihood

$$l_P(\underset{\sim}{\Theta}) = -\frac{1}{2}\log\left|\underset{\sim}{H}(\underset{\sim}{\Theta})\right| - \frac{1}{2}\left[\underset{\sim}{v} - \underset{\sim}{X}\dot{\underset{\sim}{\gamma}}(\underset{\sim}{\Theta})\right]' \underset{\sim}{H}^{-1}(\underset{\sim}{\Theta})\left[\underset{\sim}{v} - \underset{\sim}{X}\dot{\underset{\sim}{\gamma}}(\underset{\sim}{\Theta})\right].$$

This iterative procedure is repeated until the difference between successive iterate values are judged to be inconsequential.

3. EXAMPLE

3.1. Data Collection

Mite counts were collected during the summer of 2009 from two blocks of trees, taken from two different commercial 'Hass' avocado orchards (A and B) located in Carpinteria, California, USA. Trees in both orchards were planted on relatively flat terrain according to a grid system consisting of rows and columns. Eight leaves were collected from each tree with 2 leaves randomly taken from each quadrant. A total of 30 trees on a 5 x 6 grid were sampled from the block in orchard A and a total of 60 trees on a 5 x 12 grid were sampled from the block in orchard B. All sampled leaves were examined under a stereoscopic microscope and the number of live mites was counted.

3.2. Data Analysis

Fitting the count data to the model in equation (1), using the pseudo-likelihood algorithm described above, produces the parameter estimates that are shown in Table I. The estimates of the spatial scale parameter ρ indicate significant spatial correlation exists in Block B. For example, the estimated correlation between mite counts taken from adjacent trees is $\exp(-0.69) = 0.23$.

On the other hand, in Block A the estimate of ρ is zero, suggesting the counts are not spatially correlated. The estimates of the tree-to-tree variance parameter σ^2 shows there is more variability between trees in Block A than Block B, and the estimate of the parameter κ shows the counts in Block A are more overly dispersed than the counts in Block B. The estimated quadrant effects $\underset{\sim}{\gamma}$ shows that the counts in Block B have a much higher mean, and this leads to a higher overall variance as well. Tests of $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$ are rejected for both blocks, indicating unequal quadrant effects. Follow-up pair wise comparisons show that for Block A, the East and West quadrants are equally 'hot' and for Block B the West and South quadrants are equally 'hot.'

Figure 1 shows 95% prediction intervals for the estimated conditional (on $\underset{\sim}{S}$) distributions of counts based on model (1). The dots are the observed mite counts plotted against the fitted conditional mean counts where the random tree effects have been replaced by their empirical best linear unbiased predictor, $\dot{\underset{\sim}{S}}(\hat{\underset{\sim}{\Theta}})$. The solid lines are the 95% prediction intervals computed for each observation using the fitted conditional means together with the corresponding fitted negative binomial distribution. The capture rate (97.5% for Block A and 94% for Block B) and the manner in which the dots span the distance between the tolerance bands (when there are enough observations to see that) is a good reflection that the Spatial Negative Binomial GLMM fits the data well.

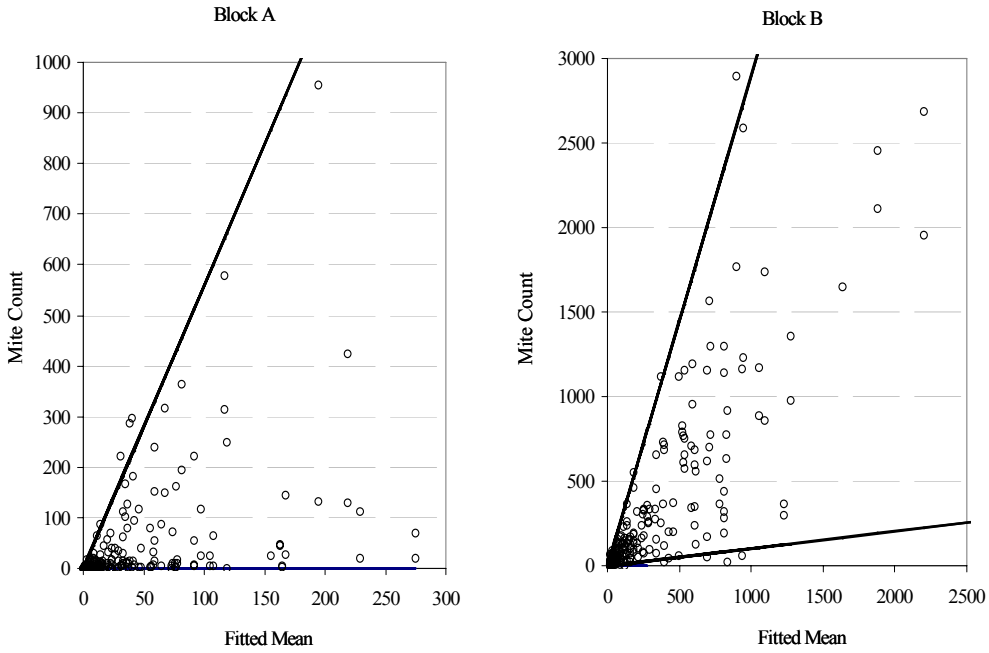

Figure 1. 95% Prediction Bands for Conditional Negative Binomial Distributions.

Table I. Fit of Spatial GLMM with Samples from Two Orchards

Parameter	Block A Estimates	Block B Estimates
ρ	5.96e-12	0.69
σ^2	2.54	1.44
κ	0.40	1.80
γ_1 (East)	2.19	5.40
γ_2 (North)	2.14	5.09
γ_3 (West)	3.24	5.25
γ_4 (South)	2.89	4.95

4. SAMPLING PLAN FOR PERIODIC PEST ASSESSMENT

Our proposed sampling plan is designed with knowledge that pest assessment is done periodically throughout each year. It is not uncommon, for example, to sample a block two or three times each year to determine if treatment is necessary. With that in mind, our proposed plan uses a fixed-size sample on the first visit to the block to estimate the magnitude of the spatial correlation in the block, and then uses that information to create a more efficient sequential sampling plan for subsequent visits to the block.

4.1. First Occasion Fixed-Size Sample

For the first occasion sampling, *m* leaves per tree are sampled from all four quadrants from adjacent trees in the middle of the block. Block sizes are typically on the order of 40×40, so sampling on the order of 5% of the trees is a good rule of thumb for getting enough information about potential spatial correlation. Sampling two leaves per quadrant will usually be sufficient to quantify the within tree variation in the pest counts. Once this data is collected, the GLMM in equation (1) can be fit, as discussed in Section 3. It is shown in that section how it can be determined if the pest associates to a 'hot' quadrant and if so, which one it is. Let γ denote the quadrant parameter corresponding to the identified 'hot' quadrant.

In terms of deciding on whether or not to treat the block, the parameter of interest is $\theta = \exp(\gamma)$, which represents the median of the conditional means associated with the GLMM in equation (1). A relevant one-sided hypothesis is $\mathrm{H}_0 : \theta \le \theta_{ET}$ vs. $\mathrm{H}_1 : \theta > \theta_{ET}$, where θ_{ET} is the economic threshold (ET) for pest counts. That is, the pest count value for which the cost of treatment is equal to the cost of incurred crop damage in the absence of treatment. Typically, θ_{ET} is the midpoint between the θ_0 and θ_1 values that were introduced in Section 1. Equivalently, the hypothesis can be expressed as $\mathrm{H}_0 : \gamma \le \log \theta_{ET}$ vs. $\mathrm{H}_1 : \gamma > \log \theta_{ET}$, which can be tested by using the standard error of $\hat{\gamma}$ to construct a upper $100(1-\alpha)\%$ confidence bound for γ. If the upper confidence bound is greater than $\log \theta_{ET}$, then $\mathrm{H}_0 : \gamma \le \log \theta_{ET}$ should be rejected and treatment should be applied to the block. Otherwise, no treatment is necessary for the block.

4.2. Subsequent Occasion Sequential Samples

The data collected from the first occasion fixed size sample the provides the information needed to develop the subsequent efficient sequential sampling schemes. In particular, the estimate of the spatial range parameter ρ allows one to determine how to sample trees to mute the effects of spatial correlation. As a rule of thumb, the so-called 'practical sampling range' is the distance at which the correlation between counts on two trees is reduced to 0.05 or less (Schabenberger and Gotway 2005). This distance can be determined once an estimate of ρ is available. When trees are sampled beyond the practical sampling range, the sampling is described as 'out of range' sampling. The first occasion sample also identifies the 'hot' quadrant.

On each subsequent occasion where a treatment decision is to be made, we propose a sequential sample of randomly selected out of range trees, taking a single leaf from the 'hot' quadrant. Sampling in this way satisfies the assumptions needed to use Bartlett's SPRT (1946) procedure for testing the simple versus simple hypothesis $\mathrm{H}_0 : \gamma = \log \theta_0$ vs. $\mathrm{H}_1 : \gamma = \log \theta_1$, which allow us to control simultaneously the type-1 and type-2 errors. For convenience, define $\gamma_0 = \log \theta_0$ and $\gamma_1 = \log \theta_1$.

4.2.1. Model Definition

Referring to equation (1), the data collected for the subsequent occasion samples will follow the reduced GLMM:

$$\begin{aligned} &a.\quad Y_i \,|\, S_i \overset{ind}{\sim} \text{Negative Binomial}(\mu_i\,,\,\kappa) \\ &b.\quad \log \mu_i = \gamma + S_i \\ &c.\quad S_i \overset{iid}{\sim} N(0,\sigma^2) \end{aligned} \tag{3}$$

where Y_i is the count recorded from the leaf sampled from the *i*-th tree, and γ is the effect of the 'hot' quadrant. Note that, unlike the model in equation (1), the S_i in equation (3) are independent since for subsequent sampling occasions the trees are sampled out of range. It can be shown that equation (3) implies that the unconditional mean and variance of Y_i are

$$\begin{aligned} \mathrm{E}(Y_i) &= \exp(\gamma + \sigma^2/2) \\ \mathrm{Var}(Y_i) &= \exp(\gamma + \sigma^2/2) + \exp(2\gamma + 2\sigma^2)(1 + 1/k) - \exp(2\gamma + \sigma^2)\,. \end{aligned}$$

4.2.2. Bartlett's SPRT

Applying Bartlett's SPRT to the observations $\{Y_1\,, Y_2\,, \dots\}$, the sequential test for subsequent sampling occasions is based on the log likelihood ratio

$$\lambda_n = \log\{ f(\underset{\sim}{Y}_n\,;\gamma_1\,,\hat{\sigma}_n^2(\gamma_1)\,,\hat{\kappa}_n(\gamma_1)) / f(\underset{\sim}{Y}_n\,;\gamma_0\,,\hat{\sigma}_n^2(\gamma_0)\,,\hat{\kappa}_n(\gamma_0))\}\,.$$

Here, n denotes the current number of observations collected in the sequential sample, $\underset{\sim}{Y}_n = (Y_1, \dots, Y_n)'$ and

$$f(\underset{\sim}{Y}_n\,;\gamma,\sigma^2,\kappa) = \prod_{i=1}^{n} \int_{s_i} \frac{\Gamma(Y_i+\kappa)}{\Gamma(Y_i+1)\Gamma(\kappa)} \left[\frac{\exp(\gamma+s_i)}{\kappa+\exp(\gamma+s_i)}\right]^{Y_i} \left[\frac{\kappa}{\kappa+\exp(\gamma+s_i)}\right]^{\kappa} \frac{1}{2\pi\sigma} e^{-s_i^2/2\sigma^2} ds_i \tag{4}$$

Equation (4) is the joint marginal distribution of the data, integrating out the random effects $\{S_i\}_{i=1}^n$. The one-dimensional integral in equation (4) is easily done using Gaussian quadrature. For $\gamma \in \{\gamma_0\,,\gamma_1\}$, the values $\hat{\sigma}_n^2(\gamma)$ and $\hat{\kappa}_n(\gamma)$ denote the conditional MLEs of the unknown nuisance parameters σ^2 and κ. These conditional MLEs are obtained by setting γ in equation (4) to γ_0 or γ_1, respectively, and then maximizing the right hand side with respect to σ^2 and κ. Define $A = \ln(\beta/(1-\alpha))$ and $B = \ln((1-\beta)/\alpha)$. The stop boundaries of Bartlett's SPRT are to accept H_0 at the first *n* for which $\lambda_n \le A$, to accept H_1 at the first *n* for which $\lambda_n \ge B$, and to continue by sampling another tree if $A < \lambda_n < B$.

The type-1 and type-2 error rates will approximately satisfy $\Pr(\text{Reject } H_0 \mid H_0) \leq \alpha$ and $\Pr(\text{Reject } H_1 \mid H_1) \leq \beta$, respectively.

4.2.3. Illustration

To illustrate, suppose a practitioner wants to test $H_0 : \theta = 3$ vs. $H_1 : \theta = 5$, or equivalently, $H_0 : \gamma = 1.1$ vs. $H_1 : \gamma = 1.6$, and that the first occasion sample determined that the out of range sampling distance is 4 trees and that the East quadrant is the 'hot' quadrant. Sampling on subsequent occasions will then be single leafs taken from the East quadrant of trees that are randomly sampled from a grid of every 5th tree. Due to the fact that Bartlett's SPRT utilizes conditional maximum likelihood estimates of the nuisance parameters σ^2 and κ, collecting an adequate number of initial samples to ensure these estimates exist and are (reasonably) reliable before initiating Bartlett's SPRT is necessary. This issue of how large the initial sample size needs to be is discussed further in the next section. For our illustration, we suppose it is decided to begin Bartlett's SPRT after an initial sample of 9 leaves has been collected. Based on nominal type-1 and type-2 errors equal to 0.1, each value of λ_n, beginning with λ_9, is compared to the stop boundaries $A = -2.2$ and $B = 2.2$. The sequential sample terminates at the first n for which either $\lambda_n \leq A$ or $\lambda_n \geq B$, and continues by sampling another leaf if $A < \lambda_n < B$.

Figure 2 depicts a hypothetical plot of λ_n for the sequence of pest counts: 5, 0, 2, 8, 2, 3, 3, 4, 10, 3, 2, 5, 4, 6, 5, 3, 0, 3, 0, 2, 0, 1. After the 9th observation was collected, Bartlett's SPRT was initiated as defined via (3) and (4), and the first value for the plot in Figure 2 corresponds to λ_9. The SPRT procedure stops at 22 trees and because $\lambda_{22} = -2.489$ is less than the lower stopping boundary $A = -2.2$. The sequential procedure terminates by accepting the null hypothesis and concluding there is no need to treat the block.

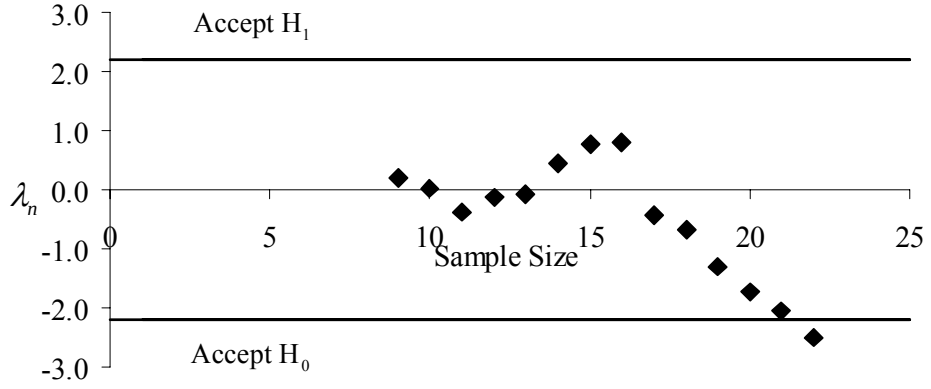

Figure 2. Hypothetical Sample Path for Bartlett's SPRT.

4.2.4. Starting Sample Size for Bartlett Test

The Operating Characteristic (OC) curve and the Average Sample Number (ASN) curve are usually used as performance criteria for a sequential procedure (Shah et al. 2009). $\text{OC}(\theta)$

is defined as the probability of accepting H_0 when θ is the true value of the parameter (Govindarajulu 2004). A desirable OC function will have $OC(\theta_0)$ and $OC(\theta_1)$ close to $1-\alpha$ and β, respectively. $ASN(\theta)$ is defined as the expected value of the number of observations required to reach a decision when θ is the true value of the parameter. It is, of course, desirable to have ASN values as small as possible, provided that $OC(\theta_0)$ and $OC(\theta_1)$ are close to their nominal values.

As mentioned above, Bartlett's SPRT has actual type-1 and type-2 error rates that are only approximately equal to the nominal values, and sufficient accuracy of the approximation depends on having a large enough number of initial samples before the Bartlett SPRT procedure formally begins. For a suitably large initial sample size, the initial wild variability in the λ_n values will be suppressed and kept from exerting an undo influence on the stopping time of the sequential procedure. One way to appreciate the importance of a suitable starting sample size is the fact that calculating λ_n requires using conditional MLEs for the nuisance parameters, and these parameters will not estimated very precisely until the number of samples collected reaches a reasonable size. For example, until the number of samples is at least equal to the number of parameters in the model, these estimates will not even exist. In this section, we report the results of a small simulation study to show the effect of the initial sample size on the realized type-1 and type-2 errors of Bartlett's SPRT.

Our simulation generates data under the model in equation (4) and uses the data analysis in Section 3.2 to guide the selection of the parameters. In particular, we choose $\sigma^2 \in \{0.5, 1.0\}$ and $\kappa \in \{1, 2, 3\}$. For each of the six combinations of (σ^2, κ), we carried out a simulation study in which we used Bartlett's SPRT to test $H_0 : E(Y) = 50$ versus $H_0 : E(Y) = 100$. (The values 50 and 100 are action thresholds currently utilized by practitioners for the persea mite.) Since $E(Y) = e^{\gamma+\sigma^2/2}$, the corresponding values of (γ_0, γ_1) depend on the value of σ^2. In particular, to achieve $E(Y \mid H_0) = 50$ and $E(Y \mid H_1) = 100$, we used $(\gamma_0, \gamma_1) = (3.66, 4.36)$ for $\sigma^2 = 0.5$, and used $(\gamma_0, \gamma_1) = (3.41, 4.11)$ for the case $\sigma^2 = 1.0$. Table II shows the mean and standard deviation for *Y* under both H_0 and H_1 for the different simulation scenarios.

Within each simulation study, we varied the initial sample size that was collected prior to formally beginning Bartlett's SPRT. In each case, 1000 sample paths were generated for each of the two cases $\gamma = \gamma_0$ and $\gamma = \gamma_1$. For each value of γ we estimated the probability of accepting H_0 by the fraction of sample paths that reached the 'Accept H_0' decision. Table III shows the initial sample size that was needed to achieve type-1 and type-2 error rates that are close to the nominal values of 0.10. We can see that as the variance in the counts gets larger (i.e., larger σ^2 and, to some extent, larger *k*) we have the intuitive result that the initial sample size needs to be larger. It seems that an initial sample size of 8-10 would be satisfactory for most pest assessment applications that utilize the model in equation (4).

Table II. Mean and Variance of Simulated Data under H_0 and H_0

σ^2	κ	H_0		H_1	
		$E(Y)$	$\sqrt{\text{Var}(Y)}$	$E(Y)$	$\sqrt{\text{Var}(Y)}$
0.5	1	50	76	100	152
	2	50	61	100	122
	3	50	55	100	110
1.0	1	50	106	100	211
	2	50	88	100	176
	3	50	81	100	162

Table III. Type-1 and Type-2 Error Rates for Alternative Starting Sample Sizes

σ^2	κ	*Initial Sample Size*	*Simulation Estimates of Error Rates*	
			Type-1	*Type-2*
0.5	*1*	*4*	*.100*	*.121*
	2	*4*	*.083*	*.112*
	3	*4*	*.063*	*.068*
1.0	*1*	*9*	*.107*	*.123*
	2	*7*	*.106*	*.116*
	3	*7*	*.106*	*.116*

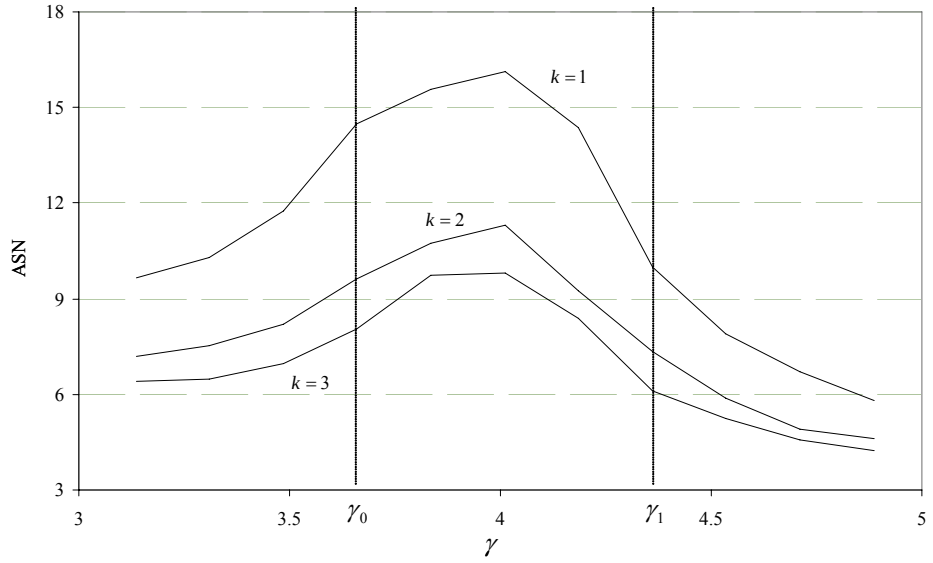

Figure 3. ASN Curve for $\sigma^2 = 0.5$ and $(\gamma_0\ ,\ \gamma_1) = (3.66\ ,\ 4.36)$.

We also used the 1000 sample paths to estimate the ASN curve for a range of 11 values of γ that include γ_0 and γ_1. Figures 3-4 show the ASN curves for the two cases $\sigma^2 = 0.5$ and $\sigma^2 = 1.0$, respectively. The figures show the intuitive result that fewer samples are

required to make a decision when the variance in the counts is less (i.e., smaller σ^2 and larger *k*).

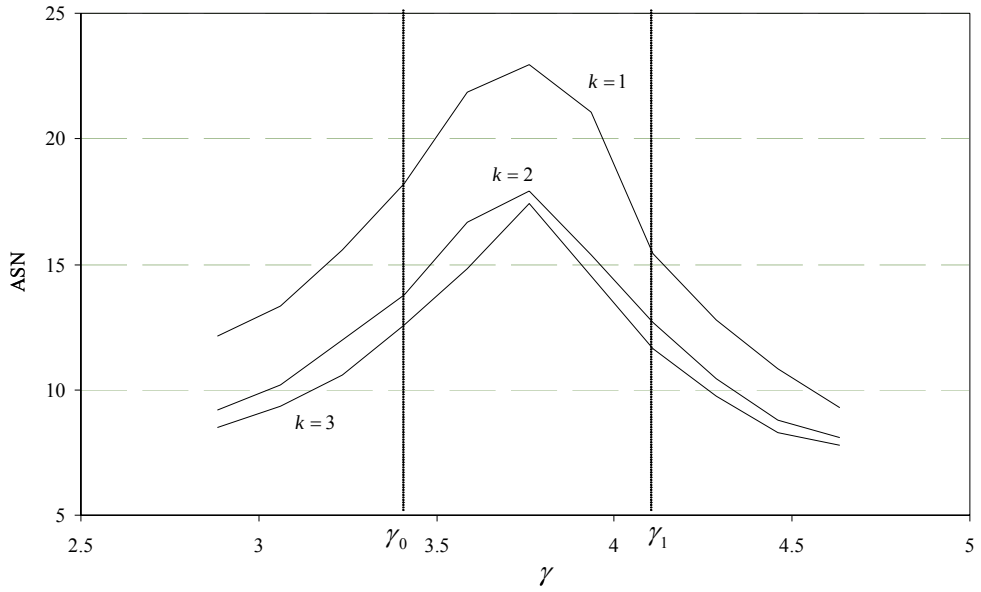

Figure 4. ASN Curve for $\sigma^2 = 1.0$ and $(\gamma_0\ ,\gamma_1) = (3.41\ ,4.11)$.

SUMMARY

We advocate the use of spatial generalized linear mixed models to analyze pest counts. Our analyses of the persea mite counts we collected demonstrate that spatial correlation in this type of data is real and can be expected. We showed how the generalized linear mixed modeling approach can be used to develop a periodic pest assessment sampling plan that features a first occasion fixed sample size and subsequent occasion sequential sample. Our proposed integration of spatial generalized linear mixed models and sequential sampling is a novel aspect of our work. The sampling plan we have proposed is practical, and the detailed description we have provided should be sufficient to enable use by pest control advisors.

REFERENCES

Barchia, I. M., G. A. Herron, and A. R. Gilmour. 2003. Use of a generalized linear mixed model to reduce excessive heterogeneity in petroleum spray oil bioassy data. *J. Econ. Entomol.* 96(3): 983-989.

Bartlett, M. S. 1946. *The large sample theory of sequential tests*. Proc. Cambridge Philos. Soc. 42: 239 – 244.

Bennett, K. E., J. E. Hopper, M. A. Stuart, M. West, and B. S. Drolet. 2008. Blood-feeding behavior of vesicular stomatitis virus infected *Culicoides sonorensis* (Diptera: Ceratopogonidae). *J. Med. Entomol*. 45: 921-926.

Bianchi E. J. J. A., P. W. Goedhart, and J. M. Baveco. 2008. Enhanced pest control in cabbage crops near forest in The Netherlands. *Landsc. Ecol.* 23: 595:602.

Binns, M. R., J. P. Nyrop, and W. Van Der Werf. 2000. *Sampling and Monitoring in Crop Protection.* CABI Publishing, New York.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. of Am. Stat. Assoc*. 88: 9-25.

Candy, S. G. 2000. The application of generalized linear mixed models to multi-level sampling for insect population monitoring. *Environ. Ecol. Stat*. 7: 217-238.

Elias, S. P., C. B. Lubelczyk, P. W. Rand, E. H. Lacombe, M. S. Holman, and R. P. Jr. Smith. 2006. Deer browse resistant exotic-invasive understory: an indicator of elevated human risk of exposure to *Ixodes scapularis* (Acari: Ixodidae) in southern coastal Maine woodlands. *J. Med. Entomol*. 43: 1142-1152.

Elston, D. A., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin. 2001. Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* 122: 563-569.

Fowler, G. W., and A. M. Lynch. 1987. Sampling plans in insect pest management based on Wald's sequential probability ratio test. *Environ. Entomol*. 16: 345-354.

Gotway, C. A., and W. W. Stroup. 1997. A generalized linear model approach to spatial data analysis and prediction. *J. of Ag. and Biol. and Environ*. Stat. 2(2) : 157-178.

Govindarajulu, Z. 2004. *Sequential statistics*. World Scientific Publishing Co. Pte. Ltd., Toh Tuck Link, Singapore.

Gozé, E., S. Nibouche, and J. P. Deguine. 2003. Spatial and probability distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in cotton: systematic sampling, exact confidence intervals and sequential test. *Environ. Entomol*. 32(5): 1203-1210.

Hoddle, M. S. 2005. Invasions of leaf feeding arthropods: Why are so many new pests attacking California-grown avocados? *California Avocado Society Yearbook* (2004-2005) 87: 65-81.

Ifoulis, A. A., and M. Savopoulou-Soultani. 2006. Use of geostatistical analysis to characterize the spatial distribution of *Lobesia botrana* (Lepidoptera: Tortricidae) larvae in northern Greece. *Environ. Entomol*. 35(2): 497-506.

Mulekar, M. S., L. J. Young, and J. H. Young. 1993. Testing insect population density relative to critical densities with 2-SPRT. *Environ. Entomol*. 22: 346-351.

Paterson, S., and J. Lello. 2003. Mixed models: getting the best use of parasitological data. *Parasitology* 19: 370-375.

Ramírez-Dávila, J. F., and E. Porcayo-Camargo. 2008. Spatial distribution of the nymphs of *Jacobiasca lybica* (Hemiptera: Cicadellidae) in a vineyard in Andalucia, Spain. Rev. *Colomb. Entomol.* 34(2): 169-175.

Schabenberger, O., and C. A. Gotway. 2005. *Statistical methods for spatial data analysis*. Chapman and Hall/CRC Press, Florida.

Schotzko, D. J., and L. E. Okeeffe. 1989. Geostatistical description of the spatial-distribution of *Lygus hesperus* (Heteroptera, Miridae) in lentils. *J. Econ. Entomol*. 82(5): 1277-1288.

Shah, P., D. R. Jeske, and R. Luck. 2009. Sequential hypothesis testing techniques for pest count models with nuisance parameters. *J. Econ. Entomol*. 102: 1970-1976.

Takakura, K. 2009. Reconsiderations on evaluating methodology of repellent effects: validation of indices and statistical analyses. *J. Econ. Entomol*. 102: 1977-1984.

Wald, A. 1947. *Sequential analysis*, Dover Publications, New York.

Wald, A., and J. Wolfowitz. 1948. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* 19: 326-339.

Williams, L., D. J. Schotzko, and J. P. Mccaffrey. 1992. Geostatistical description of the spatial distribution of *Limonius californicus* (Coleoptera, Elateridae) wireworms in the northwestern United States, with comments on sampling. *Environ. Entomol*. 21(5): 983-995.

Wolfinger, R., and O'Connell, M. 1993. Generalized linear mixed models: a Pseudo-likelihood approach. *J. of Statistical Computation and Simulation*. 48: 233-243.

Young, L. J., and J. H. Young. 1998. *Statistical ecology*. Kluwer Academic Publishers, Boston.