Co-clustering Spatial Data Using a Generalized Linear Mixed Model With Application to the Integrated Pest Management

Zhanpan Zhang, Daniel R. Jeske, Xinping Cui & Mark Hoddle

Journal of Agricultural, Biological, and Environmental Statistics

ISSN 1085-7117
Volume 17
Number 2

JABES (2012) 17:265-282
DOI 10.1007/s13253-012-0089-7

Springer

Co-clustering Spatial Data Using a Generalized Linear Mixed Model With Application to the Integrated Pest Management

Zhanpan ZHANG, Daniel R. JESKE, Xinping CUI, and Mark HODDLE

Co-clustering has been broadly applied to many domains such as bioinformatics and text mining. However, model-based spatial co-clustering has not been studied. In this paper, we develop a co-clustering method using a generalized linear mixed model for spatial data. To avoid the high computational demands associated with global optimization, we propose a heuristic optimization algorithm to search for a near optimal co-clustering. For an application pertinent to Integrated Pest Management, we combine the spatial co-clustering technique with a statistical inference method to make assessment of pest densities more accurate. We demonstrate the utility and power of our proposed pest assessment procedure through simulation studies and apply the procedure to studies of the persea mite (*Oligonychus perseae*), a pest of avocado trees, and the citricola scale (*Coccus pseudomagnoliarum*), a pest of citrus trees.

Key Words: GLMM; Heuristic optimization; Integrated pest management; Spatial co-clustering.

1. INTRODUCTION

The clustering methodology developed in this paper was motivated by a desire to improve the methodologies used by Integrated Pest Management (IPM) practitioners. IPM is an approach to managing pests by combining biological, cultural, physical and chemical tools in a way that minimizes economic losses, while simultaneously reducing human health and environmental risks. Traditional IPM practices are typically based on a hypothesis test about a parameter θ that reflects the pest density within a large block of trees in

Zhanpan Zhang is Statistician, GE Global Research, One Research Circle, Niskayuna, NY 12309, USA. Daniel R. Jeske (✉) is Professor (E-mail: *daniel.jeske@ucr.edu*) and Xinping Cui is Associate Professor, Department of Statistics, University of California, 900 University Avenue, Riverside, CA 92521, USA. Mark Hoddle is Extension Specialist in Biological Control, Department of Entomology, University of California, 900 University Avenue, Riverside, CA 92521, USA

Journal of Agricultural, Biological, and Environmental Statistics, Volume 17, Number 2, Pages 265–282
DOI: 10.1007/s13253-012-0089-7

an orchard, such as the mean or median number of pests on each tree. The hypothesis test is formulated as $H_0 : \theta \leq \theta_c$ vs. $H_a : \theta > \theta_c$, where θ_c is a critical economic threshold for which the cost of treatment is equal to the cost of no treatment. Not rejecting H_0 would indicate no treatment intervention is required, whereas rejecting H_0 would call for treatment in an attempt to ward off serious crop loss (e.g., spraying pesticides or the release of natural enemies for pest control). With the current practices, if pesticides are deemed necessary they are applied to the entire block rather than localized to smaller regions within the block. Hence, consistent with the goal of reducing unnecessary pesticide applications, a more targeted procedure that identifies smaller localized regions with high pest infestations would be useful. To meet this goal, we develop a spatial co-clustering methodology.

Co-clustering, also called biclustering, bivariate clustering, or two-mode clustering, has been broadly applied to many domains such as bioinformatics and text mining. Usually data are arranged in a matrix with rows and columns, and each cell of this matrix is a real number. Different from the one-dimensional clustering methods that seek to identify similar rows and columns independently, co-clustering seeks to take advantage of dependencies by simultaneously clustering rows and columns. Busygin, Prokopyev, and Pardalos (2008), Madeira and Oliveira (2004), Mechelen, Bock, and Boeck (2004), and Prelic et al. (2006) provided detailed reviews on co-clustering. However, there is very little literature about model-based co-clustering, and none of the literature has proposed a spatial co-clustering technique. Unlike the bioinformatics and text mining co-clustering applications, spatial co-clustering applications require that the co-clusters consist of spatially consecutive rows and columns.

The rest of this paper is organized as follows. Section 1 concludes below by introducing the application that motivated our work. In Section 2, we introduce a generalized linear mixed model (GLMM) for count data that provides for spatial correlation. We then develop a methodology to identify co-clusters from a grid sample of data that follow this GLMM. To avoid the high computational demands associated with global optimization, we propose a heuristic optimization algorithm to search for a near optimal co-clustering. The performance of the heuristic optimization algorithm is discussed. In Section 3, we incorporate the co-clustering methodology within an Integrated Pest Management (IPM) framework. Specifically, we combine the co-clustering methodology with a statistical inference procedure to propose a method for identifying regions within an orchard that need pest treatment. Finally, in Section 4 we illustrate an application of our proposed methodology using studies of the persea mite (*Oligonychus perseae*), a pest of avocado trees, and the citricola scale (*Coccus pseudomagnoliarum*), a pest of citrus trees.

2. METHODOLOGY

2.1. GLMM FOR CO-CLUSTERING

2.1.1. Model Definition

Consider an $r \times c$ spatial grid in which each grid point is a potential sampling site. By simultaneously dividing rows and columns into a number of contiguous and disjoint groups, we obtain a checkerboard structure within the grid. Each of the groups is referred to

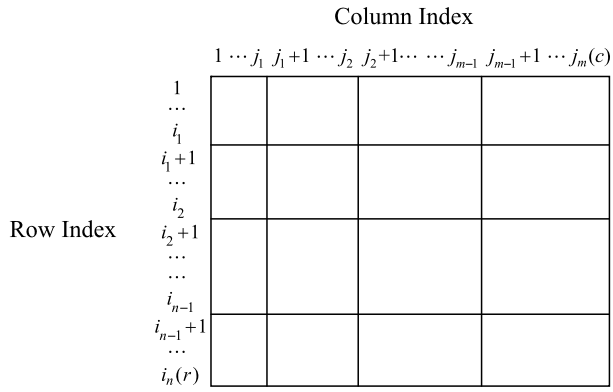


Figure 1. Checkerboard co-cluster structure.

as a co-cluster and the ensemble of co-clusters is called a co-clustering. For a co-clustering that has n groups of rows and m groups of columns, we use the term "nomenclature" to refer to the $n \times m$ checkerboard structure. We use the term "design" to refer to the specific rows and columns within a given nomenclature. Referring to Figure 1, a specific design within an $n \times m$ nomenclature is denoted by $(i_1, i_2 - i_1, \ldots, i_n - i_{n-1}) \times (j_1, j_2 - j_1, \ldots, j_m - j_{m-1})$. Note that there exists a many-to-one mapping between designs and nomenclatures.

The GLMM we propose for co-clustering according to a checkerboard structure is

$$\begin{aligned} Y_{j(i)}|\mathbf{s} &\overset{ind}{\sim} \text{Negative Binomial}(\theta_i, \kappa), \quad i = 1, 2, \ldots, nm, \, j = 1, 2, \ldots, n_i; \\ \log(\theta_i) &= \mu + s_i; \\ \mathbf{s} = (s_1, s_2, \ldots, s_{nm})' &\sim \text{MVN}\big(0, \sigma^2 \mathbf{I}_{nm}\big); \end{aligned} \tag{2.1}$$

where $Y_{j(i)}$ is the count number from the jth sampling unit in the ith co-cluster, n_i is the number of sampling units in the ith co-cluster, θ_i is the conditional (on s_i) mean of counts associated with the ith co-cluster, κ represents an overdispersion parameter, μ is a fixed intercept effect, and s_i is a random effect associated with the ith co-cluster, and $\mathbf{I}_{nm}$ is the identity matrix of size nm.

2.1.2. Model Justification for IPM Application

Recent literature has shown that pest density levels are influenced by spatial population dynamics. For example, spatial analyses have been applied in studies of agricultural pests attacking lentils (Schotzko and O'Keeffe 1989), corn and alfalfa (Williams, Schotzko, and McCaffrey 1992), cotton (Gozé, Nibouche, and Deguine 2003), and grapes (Ifoulis and Savopoulou-Soultani 2006; Ramírez-Dávila and Porcayo-Camargo 2008). Typically, spatial analyses of pest density populations have been conducted by transforming pest counts to approximately satisfy the normality assumption. GLMMs (Breslow and Clayton 1993) are a more natural way to describe pest counts, and through the correlations introduced by their random effects they can be used to model various types of spatial correlation. Our use of GLMMs to develop a model-based spatial co-

clustering methodology differs from a variety of ways GLMMs have previously been used in studies of pest populations (Barchia, Herron, and Gilmour 2003; Bennett et al. 2008; Bianchi, Goedhart, and Baveco 2008; Candy 2000; Elias et al. 2006; Elston et al. 2001; Paterson and Lello 2003; Takakura 2009).

Consider now the independence assumption in (2.1) for the random effects. Insects who are in search of food tend to invade orchards by establishing small populations in random areas. The haphazard and often diffuse initial settlements of high density populations in small areas are referred to by pest management specialists as "hot spots." IPM procedures are carried out before or at the outbreak of hot spots, before they blend together to form larger clusters with smoother densities, by which time economic damage in the orchard will be so severe that mitigation measures will be ineffective. The existence of hot spots is more compatible with the checkerboard spatial correlation structure implied by (2.1) than it would be with smooth structures often seen in other spatial applications (e.g., Gotway and Stroup 1997).

2.1.3. Likelihood and Parameter Estimation

It follows from standard GLMM principles (McCulloch, Searle, and Neuhaus 2008) that the log-likelihood corresponding to (2.1) is

$$
\begin{aligned}
l(\mu,\sigma^2,\kappa) &= \sum_{i=1}^{nm}\Bigg[\log\int_{-\infty}^{\infty}\Bigg(\prod_{j=1}^{n_i}\Bigg(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\Bigg)\Bigg(\frac{\kappa}{\exp(\mu+s_i)+\kappa}\Bigg)^{\kappa} \\
&\quad\times\Bigg(\frac{\exp(\mu+s_i)}{\exp(\mu+s_i)+\kappa}\Bigg)^{y_{j(i)}}\Bigg)\cdot\frac{\exp(-s_i^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}}ds_i\Bigg] \\
&= \sum_{i=1}^{nm}\sum_{j=1}^{n_i}\log\Bigg(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\Bigg)+\sum_{i=1}^{nm}\log\Bigg[\int_{-\infty}^{\infty}\Bigg(\frac{\kappa}{\exp(\mu+s_i)+\kappa}\Bigg)^{n_i\kappa} \\
&\quad\times\Bigg(\frac{\exp(\mu+s_i)}{\exp(\mu+s_i)+\kappa}\Bigg)^{\sum_{j=1}^{n_i}y_{j(i)}}\frac{\exp(-s_i^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}}ds_i\Bigg]. \qquad (2.2)
\end{aligned}
$$

Equation (2.2) involves nm one-dimensional integrals, each of which can be approximated as a weighted sum by the method of Gauss–Hermite quadrature:

$$
\begin{aligned}
l(\mu,\sigma^2,\kappa) &\approx \sum_{i=1}^{nm}\sum_{j=1}^{n_i}\log\Bigg(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\Bigg) \\
&\quad+\sum_{i=1}^{nm}\log\Bigg[\sum_{d=1}^{D}\Bigg(\Bigg(\frac{\kappa}{\exp(\mu+\sqrt{2\sigma^2}x_d)+\kappa}\Bigg)^{n_i\kappa} \\
&\quad\times\Bigg(\frac{\exp(\mu+\sqrt{2\sigma^2}x_d)}{\exp(\mu+\sqrt{2\sigma^2}x_d)+\kappa}\Bigg)^{\sum_{j=1}^{n_i}y_{j(i)}}\frac{w_d}{\sqrt{\pi}}\Bigg)\Bigg], \qquad (2.3)
\end{aligned}
$$

where x_d's and w_d's ($d=1,2,\ldots,D$) are the quadrature nodes and weights, respectively. Quadrature with $D=30$ is usually enough for a good degree of approximation (McCul-

loch, Searle, and Neuhaus 2008), and we use that in all of the calculations in this paper. Then (2.3) can be maximized numerically to obtain the MLEs of (μ, σ^2, κ), denoted as $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.

2.2. CO-CLUSTERING ALGORITHMS

2.2.1. Global Optimization

We define the optimal design to be the one with the maximum log-likelihood among all the possible designs. To avoid co-clusters that are too small, we specify the minimum co-cluster size to be $r_0 \times c_0$ ($r_0 > 1$ and $c_0 > 1$). A global optimization algorithm (GOA) would identify all possible designs for every possible nomenclature, and select the global optimal design as the one that maximizes $l(\mu, \sigma^2, \kappa)$.

Zhang (2011) shows that the number of designs that need to be examined when searching for the global optimal design is

$$\left[1+\sum_{n=2}^{\lfloor r/r_0\rfloor}\left\{\binom{r-1}{n-1}+\sum_{i=1}^{n-1}\left[(-1)^i\binom{n}{i}\sum_{j_1,j_2,\ldots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}\right]\right\}\right]$$
$$\times\left[1+\sum_{m=2}^{\lfloor c/c_0\rfloor}\left\{\binom{c-1}{m-1}+\sum_{i=1}^{m-1}\left[(-1)^i\binom{m}{i}\sum_{j_1,j_2,\ldots,j_i=1}^{c_0-1}\binom{c-(j_1+j_2+\cdots+j_i)-1}{m-i-1}\right]\right\}\right]. \quad (2.4)$$

To illustrate, for a spatial grid of size 80×80 and a minimum co-cluster size of 12×12, the number of possible designs is 382,241,601. The enormous number of candidates usually makes it infeasible to exhaustively search for the optimal design.

2.2.2. Heuristic Optimization

To circumvent the computational complexity associated with global optimization, we propose the following heuristic optimization algorithm (HOA):

(1) Starting with the original spatial grid, fit the GLMM to each of the designs associated with the 1×2 and 2×1 nomenclatures. Identify the design with the maximum $l(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ as the "Current Optimal Design" and denote its log-likelihood by $l^*(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.

(2) Starting with the "Current Optimal Design", fit the GLMM to each of the designs with the nomenclature that has either one more row group or one more column group than the "Current Optimal Design." Identify the design with the maximum $l(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ as the "Potential Optimal Design" and denote its log-likelihood is by $l^0(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.

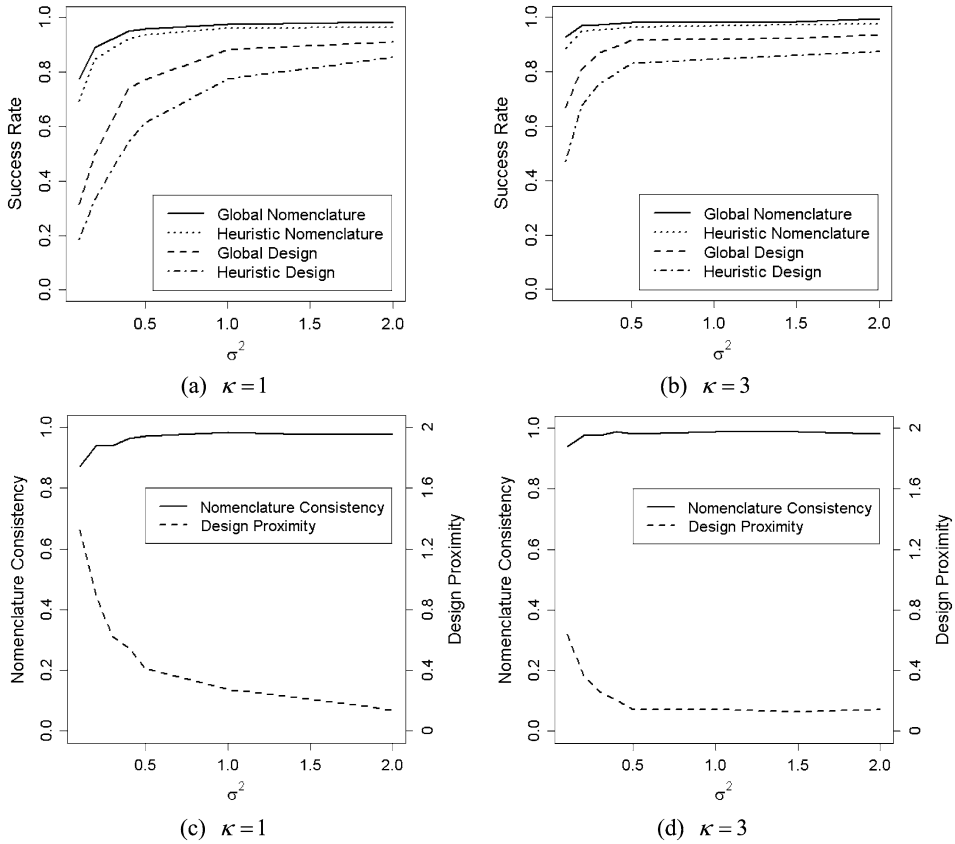


Figure 2. Heuristic optimization vs. global optimization.

(3) If $l^0(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa}) > l^*(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$, replace the "Current Optimal Design" with the "Potential Optimal Design" and repeat Step 2; otherwise, stop the procedure and report the "Current Optimal Design" as the heuristic optimal design.

2.3. Comparative Analysis

2.3.1. Heuristic vs. Global Optimization

To study the effectiveness of our proposed HOA, we performed a simulation study to compare it to the GOA according to Steps 1–5 in Simulation Design 1 shown in Table A.1 of Appendix A. Results are summarized in Figures 2(a) and 2(b) which show the success rates for the reported nomenclature and design for different (κ, σ^2) scenarios.

In this simulation study, the number of designs evaluated in the global optimization algorithm is 2937, whereas the average number of designs evaluated in the heuristic optimization algorithm is 85. From either Figure 2(a) or 2(b), we see the success rate of the nomenclature or design for the HOA is not that much lower than that for the GOA. Also, the success rates increase as σ^2 increases, which indicates that greater difference among true co-clusters improves the chance of retrieving the true design or nomenclature. We also see

that success rates increase as κ increases, meaning that less conditional variability within true co-clusters also improves the chance of capturing the true design or nomenclature.

To further compare HOA with GOA, we define "nomenclature consistency" to be the proportion of times that the two algorithms report the same nomenclature (which may or may not be same as the true nomenclature). Given that the two algorithms report the same nomenclature, the number of shifts in the row and column separation lines that are needed in order to match the two reported designs can then be used as a measure of "design proximity." To study nomenclature consistency and design proximity, we performed the simulation study described by steps (1–3, 4′–5′) in Simulation Design 1 that is shown in Table A.1 of Appendix A. Figures 2(c) and 2(d) shows results for different (κ, σ^2) scenarios.

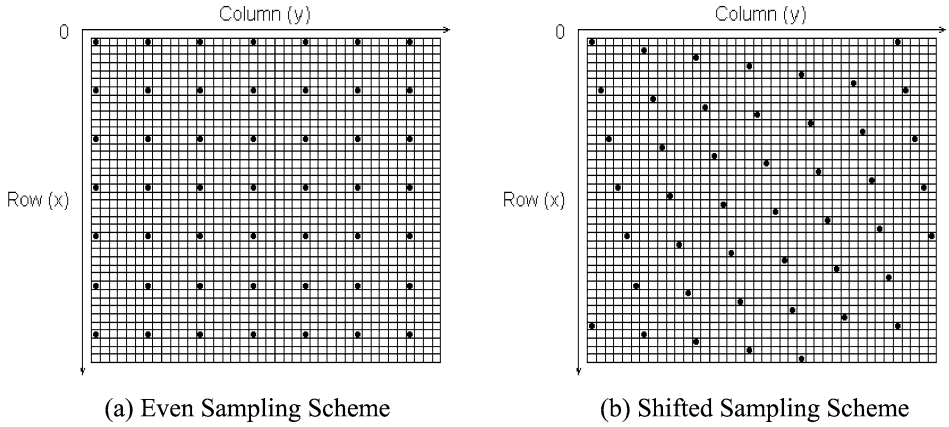
From either Figure 2(c) or 2(d) we notice the nomenclature consistency is very high, meaning that very often both the GOA and HOA report the same nomenclature. In addition, given the same reported nomenclature, the average design proximity is very small (less than 2). These results demonstrate the HOA can perform very well as a surrogate for the GOA.

2.3.2. Sampling Issues

Concerning time and the cost of human resources, practitioners usually do not exhaustively sample the full set of grid points. If simple random sampling is used, it is very likely that specific areas of the spatial grid will not be represented in the sample. In this case, we can anticipate that some of the resulting co-clusters will not have been sampled and in some applications, such as the one we discuss in Section 3, this can lead to loss of precision in subsequent inference procedures.

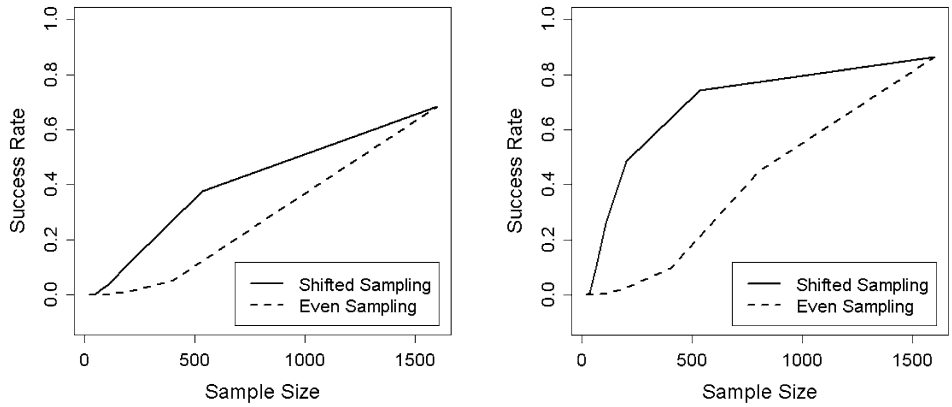
Recall that the minimum co-cluster size is $r_0 \times c_0$. To ensure at least one grid point is taken from each co-cluster, we could start by sampling the grid point located in the first row and the first column, and then sample a grid point every r_0 rows along the row dimension and every c_0 columns along the column dimension. This strategy is illustrated in Figure 3(a) for the case 40×40 with a minimum co-cluster size of 6×6. Note, however, two designs that differ in terms of where their column and row separation lines are positioned but otherwise have the same configuration of sampled points in their co-clusters will have identical likelihood values.

To minimize the number of co-cluster configurations that have identical likelihood values, we propose the sampling strategy shown in Figure 3(b) in which each sampled grid point is shifted one more row than the previously sampled grid point along the column dimension, and shifted one more column than the previously sampled grid point along the row dimension. By following this shifted pattern, a few more grid points will be sampled from the top-right and/or bottom-left corner of the spatial grid to ensure at least one grid point is sampled from each co-cluster. Compared to the even sampling scheme, without significantly increasing the sample size the shifted sampling scheme minimizes the number of rows and columns that are not represented by at least one sampled grid point. Consequently, the shifted scheme reduces the number of designs that have identical likelihood values and will have more discriminatory design power than the even sampling scheme.



(a) Even Sampling Scheme (b) Shifted Sampling Scheme

Figure 3. Non-exhaustive sampling strategy.



(a) $\sigma^2 = 0.2$ (b) $\sigma^2 = 1$

Figure 4. Success rate of design vs. sample size.

For a specified minimum co-cluster size, the proposed sampling strategy seeks to optimize the discriminatory design power and minimize the sample size subject to the constraint that at least one grid point is sampled from each co-cluster. When practitioners can afford to sample more grid points, the sample size can be increased by replacing r_0 with a smaller "row step" r^* $(1 \le r^* < r_0)$ and c_0 with a smaller "column step" c^* $(1 \le c^* < c_0)$ such that any $r^* \times c^*$ sub-grid contains at least one sampled grid point. For example, $r^* = c^* = 4$ leads to 108 sampled grid points in Figure 3(b) for the shifted sampling strategy, as compared to 46 sampled grid points if no reduction in row and column steps was implemented.

The simulation study described by Simulation Design 2 (shown in Table A.2 of Appendix A) was performed to evaluate how non-exhaustive sample sizes affect the success rate of the design. The results are summarized in Figure 4 which shows the relationship between success rate of the design and the sample size. From either Figure 4(a) or 4(b), we see the success rate of the design increases as the sample size increases and the success

rate of the design for the shifted sampling strategy is much higher than that for the even sampling strategy. Also, similar to what we saw in Figure 2, we see that the success rate increases as σ^2 increases.

3. APPLICATION TO PEST DENSITY ASSESSMENT

3.1. PROPOSED METHODOLOGY

We consider an application to assess orchards of fruit-bearing trees for a potential pest problem. Our goal is to identify the infested regions within orchards that require treatment such as spraying pesticides or, alternatively, the release of natural enemies. Trees within orchards are frequently organized in a spatial grid. We assume that pest counts are available from a sample of trees based on the shifted sampling strategy described in Section 2.3.2. Our proposed methodology consists of first using the HOA on the sampled data to obtain the heuristic optimal design of the orchard. Then, in a second step we analyze the data further, as described in the next section, to make a decision on whether pest treatment intervention is required.

3.2. INFERENCE-BASED TREATMENT DECISIONS

For each co-cluster that results from the HOA, we use the model in (2.1) to predict its conditional mean $\theta_i = \exp(\mu + s_i)$ $(i = 1, 2, \ldots, nm)$. It is shown in Appendix B that the Best Linear Predictor (BLP) is

$$\begin{aligned}\tilde{\theta}_i &= \mathrm{BLP}(\theta_i) \\ &= \frac{\exp(\mu+\sigma^2/2)(\exp(\sigma^2)-1)\cdot\sum_{j=1}^{n_i} y_{j(i)} + \exp(2\mu+2\sigma^2)/\kappa + \exp(\mu+\sigma^2/2)}{\exp(\mu+3\sigma^2/2)/\kappa + 1 + n_i\exp(\mu+\sigma^2/2)(\exp(\sigma^2)-1)},\end{aligned} \tag{3.1}$$

and that the variance of $\log\tilde{\theta}_i - \log\theta_i$ is

$$\begin{aligned}\mathrm{Var}(\log\tilde{\theta}_i - \log\theta_i) \approx \big\{&n_i\exp\big(2\mu+\sigma^2\big)\big(\exp\big(\sigma^2\big)-1\big)^2\big[\exp\big(\sigma^2/2-\mu\big)+1/\kappa\big] \\ &+\exp\big(2\sigma^2\big)\big(\exp\big(\sigma^2\big)-1\big)\big[\exp\big(\mu+3\sigma^2/2\big)/\kappa+1\big]^2\big\} \\ &\div\big[\exp\big(\mu+3\sigma^2/2\big)/\kappa+1+n_i\exp\big(\mu+\sigma^2/2\big)\big(\exp\big(\sigma^2\big)-1\big)\big]^2.\end{aligned} \tag{3.2}$$

Replacing (μ, σ^2, κ) with the MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ in (3.1) and (3.2) gives the empirical Best Linear Predictor (eBLP), say $\hat{\tilde{\theta}}_i$, and the estimated variance of $\log\tilde{\theta}_i - \log\theta_i$, say $\widehat{\mathrm{Var}}(\log\tilde{\theta}_i - \log\theta_i)$.

Define $U_i = (\log\hat{\tilde{\theta}}_i - \log\theta_i)/\sqrt{\widehat{\mathrm{Var}}(\log\tilde{\theta}_i - \log\theta_i)}$ and let $U_{i,\alpha}$ be the $100(1-\alpha)$th conditional percentile of U_i given $\mathbf{s}$. Then a $100(1-\alpha)$ % lower conditional prediction bound for $\log\theta_i$ is

$$L_\alpha(\log\theta_i) = \log\hat{\tilde{\theta}}_i - U_{i,\alpha}\sqrt{\widehat{\mathrm{Var}}(\log\tilde{\theta}_i - \log\theta_i)}, \tag{3.3}$$

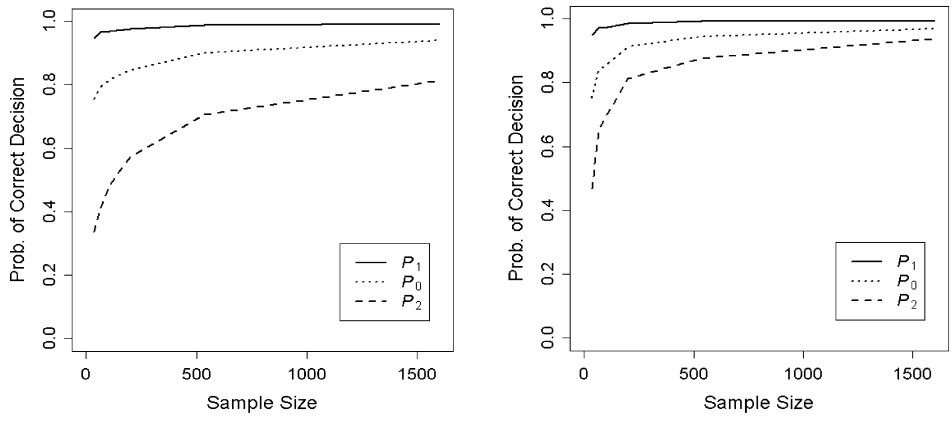


Figure 5. Probabilities of correct decision.

and a $100(1-\alpha)$ % lower conditional prediction bound for θ_i is

$$L_\alpha(\theta_i) = \hat{\tilde{\theta}}_i / \exp\left[U_{i,\alpha}\sqrt{\widehat{\mathrm{Var}}(\log\tilde{\theta}_i - \log\theta_i)}\right]. \tag{3.4}$$

For a pre-specified threshold θ_c, the decision of "Treat" is made if $L_\alpha(\theta_i) > \theta_c$; otherwise the decision of "Do Not Treat" is made. The value of $U_{i,\alpha}$ in (3.3) and (3.4) can be approximated from the parametric bootstrap algorithm described in Appendix C.

A simulation study was conducted to verify the coverage probabilities based on the bootstrap estimates of $U_{i,\alpha}$ were satisfactorily close to nominal. When the number of co-clusters of the heuristic optimal design is relatively large, we may adjust the significance level α to form the simultaneous lower conditional prediction bounds of the conditional means for co-clusters, such as by the method of Bonferroni correction or Sidak correction (Olejnik et al. 1997).

3.3. Performance Analysis

To evaluate the proposed pest assessment procedure outlined in Section 3.1, we performed a simulation study according to Simulation Design 3 that is shown in Table A.3 of Appendix A. We selected $\alpha = 0.05$ and chose the critical economic threshold of $\theta_c = 500$ based on the analyses discussed in Section 4.1. The simulation provides estimates of the conditional probabilities $P_1 = P(\text{Correct decision}|\text{Truth is "Do Not Treat"})$, $P_2 = P(\text{Correct decision}|\text{Truth is "Treat"})$, and the unconditional probability $P_0 = P(\text{Correct decision})$. The results are summarized in Figure 5, from which we notice all of P_0, P_1 and P_2 increase as the sample size increases, and also increase as σ^2 increases. The relatively high values, for moderate sample size and σ^2, demonstrate the practical utility of our proposed pest assessment procedure.

4. ILLUSTRATION

4.1. PERSEA MITES AND AVOCADO TREES

Persea mite (*Oligonychus perseae*) 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). When pest populations build to sufficiently high densities leaves begin to drop from trees. To avoid premature leaf dropping some type of control procedure may be warranted (e.g., pesticide applications, or releases of commercially available natural enemies, like predatory mites that eat the pest).

A data set of mite counts is available from a pilot monitoring study conducted in the Summer of 2009 at a large commercial avocado orchard located near Carpenteria, California, USA. Trees in this orchard are planted on a large grid structure. The available data were obtained by sampling all the trees on a small 5×12 grid. Eight leaves were collected from each tree, and their sum provided a pest count for each sampled tree.

Using a minimum co-cluster size of $r_0 \times c_0 = 2 \times 3$ and applying the HOA, we obtained the heuristic optimal design as shown in Figure 6(a), in which four co-clusters are separated by the thick dark column separations. The 95 % lower conditional prediction bound of the conditional mean for each co-cluster is positioned at the top of each co-cluster. Using a critical threshold of $\theta_c = 500$ (see Maoz et al. 2011), it is clear that three of the co-clusters require treatment.

4.2. CITRICOLA SCALES AND CITRUS TREES

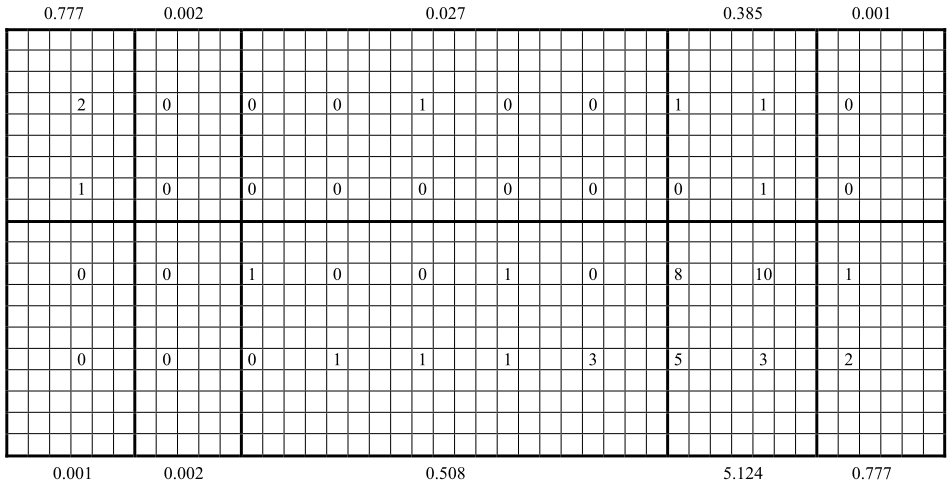
In the previous illustration, exhaustive sampling was carried out over a relatively small region within the orchard. This illustration reflects non-exhaustive sampling, which is more typical within large orchards. The citricola scale is a pest of citrus in California's San Joaquin Valley. They are almost microscopic in size and have a one year life span. Each female citricola scale can lay between 1000 and 1500 eggs after which its body covers and protects the eggs until they hatch. A study in Shah (2006) presents citricola samples that were collected from blocks of orange trees that were laid out on a spatial grid. In that study, individual blocks were considered as treatable units. We reconsider those data to illustrate how our co-clustering methodology identifies more focused hot spots within a block.

Two of the blocks sampled by Shah (2006) were laid out on a spatial grid. These were blocks sampled from the Rolling Hills and Porterville orchards. Figures 6(b) and 6(c) show these blocks, where the cells with numbers in them are the locations of sampled trees. From each of the sampled tree, approximately equal-sized twigs were drawn from the bottom half of the tree, facing north, based on a-priori information that the citricola scale has a slight preference for this part of the tree. The number of scales found on each sampled twig was then counted and these data are the numbers shown in Figures 6(b) and 6(c).

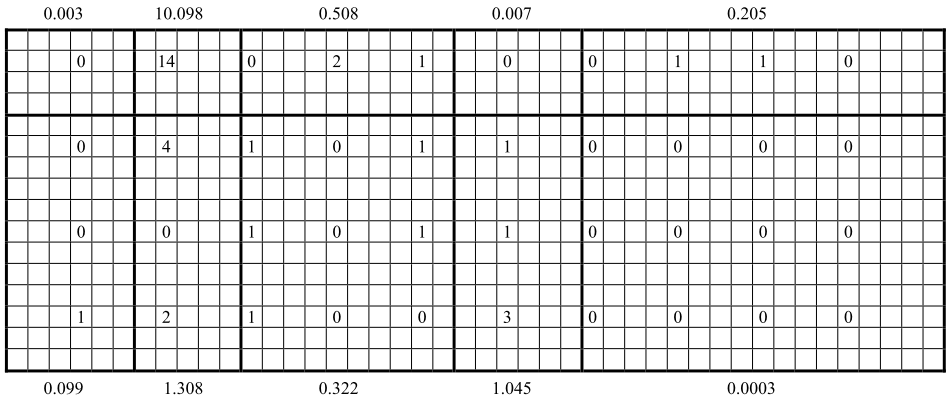
In our analysis, we used a minimum co-cluster size of $r_0 \times c_0 = 5 \times 5$ for the Rolling Hills block and $r_0 \times c_0 = 4 \times 5$ for the Porterville block. The GOA would require evaluating 994,296 and 760,344 designs, respectively, for these two cases. The HOA produced nomenclatures of size 2×5 for both blocks, as indicated by thick dark row and column

1700			309			3196			1465		
1154	332	3318	388	187	415	8259	791	1847	1554	3595	3403
3632	1288	481	120	121	958	5647	1075	1513	1994	628	1066
3221	1259	5261	599	1182	172	3621	8592	5431	1566	2052	5982
1519	843	3577	517	307	177	18344	2365	6828	336	4411	3515
6371	1730	5592	254	228	5288	2604	452	3856	693	2899	408

(a) Counts of persea mites at 60 avocado trees laid out as a 5×12 spatial grid in a Carpenteria, CA orchard. Numbers adjacent to the identified co-clusters are the 95% lower conditional prediction bounds for the conditional means.



(b) Counts of citricola scales at 40 orange trees from a block laid out as a 20×44 spatial grid in the Rolling Hills, CA orchard. Numbers adjacent to the identified co-clusters are the 95% lower conditional prediction bounds for the conditional means.



(c) Counts of citricola scales at 40 orange trees from a block laid out as a 16×44 spatial grid in the Porterville, CA orchard. Numbers adjacent to the identified co-clusters are the 95% lower conditional prediction bounds for the conditional means.

Figure 6. Pest treatment decision.

separations. The number of designs examined by the heuristic algorithm was 164 and 169, respectively. As discussed in Section 2.3.2, if different designs are obtained by moving the nomenclature boundaries without changing the co-cluster membership of sampled trees, the designs will have identical likelihood values. We can see in Figures 6(b) and 6(c) that due to the even sampling strategy used by Shah (2006), there are several likelihood-equivalent designs returned by the HOA.

Also in Figures 6(b) and 6(c), we have positioned the 95 % lower conditional prediction bound next to each co-cluster. To-date, a definitive critical economic threshold, θ_c, for the number of Citricola scales on a sampled twig is not well known. However, it is known that the tolerance for early appearance of the insect in the San Joaquin Valley is extremely low due to the fact there are no natural enemies of the insect in that region of California. Consequently, for illustrative purposes suppose 0.5 is used as the threshold. It would follow that four co-clusters in each of the blocks in Figures 6(b) and 6(c) would be deemed as in need of treatment.

As point of comparison, we also ran the GOA on the data from the Rolling Hills and Porterville blocks. In the Rolling Hills case, the result was the same 2×5 nomenclature obtained from the HOA, and the design was also the same up to small perturbations in the column and row separations that have no affect on the log-likelihood value (whose value was -45.68 for both algorithms).

In the Porterville case, the GOA returned a 2×6 nomenclature, compared to the 2×5 nomenclature returned by the HOA. Examining Figure 6(c), an additional column separator was added to isolate the third column of samples from the fourth and fifth columns of samples. The co-cluster identified with a 95 % lower conditional prediction bound of 0.508 is split into two co-clusters with bounds of 0.006 and 0.895, and the co-cluster with a 95 % lower conditional prediction bound of 0.322 is split into two co-clusters with bounds of 0.548 and 0.145. These are relatively minor changes, and this is reflected by the fact the log-likelihood value for the HOA is -43.33 compared to -42.87 for the GOA.

5. DISCUSSION

Our proposed model-based co-clustering method showed a significant utility and power in searching for the optimal co-clustering on a spatial grid. Combining the spatial co-clustering technique with a statistical inference method, our proposed pest assessment procedure also showed an excellent performance in identifying the infested regions within orchards. Only treating the infested regions instead of the whole orchard can reduce pest management costs and minimize potential hazards to the environment. Although these methods were developed to analyze the pest data collected from perennial tree orchards, we anticipate that this general approach will have utility for a wide range of investigations involving spatial information.

In this paper we used a GLMM to capture correlation within co-clusters, and as discussed in Section 2.1.2, assumed all the co-clusters to be independent of each other. Although this assumption makes much practical sense with our application, we will further

consider a GLMM that captures both correlation within co-clusters and correlation between co-clusters as future work. Furthermore, more flexible co-cluster structures will be investigated for the spatial grid in future work, such as the "tree" co-cluster structure that was considered in Hartigan (1972), which is considered the first co-clustering paper.

APPENDIX A: SIMULATION DESIGNS

Parameter values used by the simulation designs discussed in this appendix were motivated by the avocado application involving persea mites that is discussed in Section 4.1.

Table A.1. Simulation Design 1.

1. Consider a 40×40 spatial grid, a 3×3 nomenclature and the $(10, 17, 13) \times (13, 15, 12)$ design. Choose the minimum co-cluster size to be $r_0 \times c_0 = 10 \times 12$.
2. For $\mu = 6$ and selected values of (κ, σ^2), simulate count data for each point on the spatial grid according to the negative binomial model and the design selected in Step 1.
3. Apply both the GOA and the HOA to the simulated data.
4. For each algorithm, check whether the true nomenclature and design that were specified in Step 1 were retrieved.
5. Repeat Steps 2–4 1000 times, and for each algorithm, record the success rates for the reported optimal design and nomenclature.

4′. Determine whether the two algorithms agree on the nomenclature. If the two nomenclatures are the same, evaluate the design proximity metric.

5′. Repeat Steps 2–4′ 1000 times and evaluate the proportion of times the two nomenclatures agree and the average design proximity.

Table A.2. Simulation Design 2.

1. Same as Step 1 in Table A.1.
2. For $(\mu, \kappa) = (6, 3)$ and a selected value of σ^2, simulate count data for each point on the spatial grid according to the negative binomial model and the design selected in Step 1.
3. For selected values of (r^*, c^*) (to vary the sample sizes), sample grid points according to both the even and shifted sampling patterns.
4. Apply the HOA to the sampled data.
5. Check whether the true design that was specified in Step 1 was retrieved.
6. Repeat Steps 2–5 1000 times and evaluate the success rate of the retrieved design.

Table A.3. Simulation Design 3.

1. Same as Step 1 in Table A.1.
2. For $(\mu, \kappa) = (6, 3)$ and a selected value of σ^2, simulate count data for each tree in the orchard according to the negative binomial model and the design selected in Step 1. Record the conditional means (θ_i's) of the nine true co-clusters.
3. For selected values of (r^*, c^*) (to vary the sample sizes), take a sample of trees according to the shifted sampling pattern.
4. Apply the HOA to the sampled data.
5. Calculate $L_\alpha(\theta_i)$ for each co-cluster of the heuristic optimal design ($i = 1, 2, \ldots, M$), where M is the number of co-clusters of the heuristic optimal design that may be different from the nine true co-clusters.
6. By comparing the conditional means (θ_i's) of the true co-clusters with $\theta_c = 500$, determine the true treatment status of trees within the true co-clusters, saying "Treat" if $\theta_i > \theta_c$, and "Do Not Treat" otherwise.
7. By comparing the $L_{.05}(\theta_i)$'s ($i = 1, 2, \ldots, M$) with $\theta_c = 500$, determine the decided treatment status of trees within the i^{th} co-cluster of the heuristic optimal design, saying "Treat" if $L_{.05}(\theta_i) > \theta_c$, and "Do Not Treat" otherwise.
8. Evaluate each tree with respect to its true treatment status and its decided treatment status and assign it into the appropriate cell of a 2×2 confusion matrix.
9. Repeat Steps 2–8 1000 times and accumulate all the comparisons of Step 8 into a pooled 2×2 misclassification table.

APPENDIX B: DERIVATION OF $\tilde{\theta}_i = \text{BLP}(\theta_i)$ AND $\text{Var}(\log \tilde{\theta}_i - \log \theta_i)$

For $j = 1, 2, \ldots, n_i$, we have

$$\text{Cov}\big(\exp(\mu + s_i), \mathbf{y}_i\big) = \exp\big(2\mu + \sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big) \cdot \mathbf{1}'_{n_i}, \tag{A1}$$

where $\mathbf{y}_i = (y_{1(i)}, y_{2(i)}, \ldots, y_{n_i(i)})'$, and $\mathbf{1}_{n_i}$ is the n_i-tuple column vector of all 1's.

For $j, j' = 1, 2, \ldots, n_i$ and $j \neq j'$, we have

$$\text{Var}(y_{j(i)}) = \exp\big(2\mu + 2\sigma^2\big)/\kappa + \exp\big(\mu + \sigma^2/2\big) + \exp\big(2\mu + \sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big),$$

$$\text{Cov}(y_{j(i)}, y_{j'(i)}) = \exp\big(2\mu + \sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big).$$

Thus $\text{Var}(\mathbf{y}_i) = (\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)) \cdot \mathbf{I}_{n_i} + \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \cdot \mathbf{J}_{n_i}$, and

$$\text{Var}^{-1}(\mathbf{y}_i) = \frac{1}{\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)} \cdot \left(\mathbf{I}_{n_i} - \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)} \mathbf{J}_{n_i}\right), \tag{A2}$$

where $\mathbf{I}_{n_i}$ is the identity matrix of size n_i and $\mathbf{J}_{n_i}$ is the n_i-by-n_i matrix with all 1's.

Hence, from (A1) and (A2), we have

$$\begin{aligned}\tilde{\theta}_i &= E(\theta_i) + \mathrm{Cov}(\theta_i, \mathbf{y}_i) \cdot \mathrm{Var}^{-1}(\mathbf{y}_i) \cdot \big(\mathbf{y}_i - E(\mathbf{y}_i)\big) \\ &= \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1) \cdot \sum_{j=1}^{n_i} y_{j(i)} + \exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.\end{aligned}$$

The first order of Taylor expansion of $\log\tilde{\theta}_i$ around θ_i gives

$$\log\tilde{\theta}_i \approx \log\theta_i + (\tilde{\theta}_i - \theta_i)/\theta_i.$$

Hence

$$\begin{aligned}\mathrm{Var}(\log\tilde{\theta}_i - \log\theta_i) &\approx \mathrm{Var}\big[(\tilde{\theta}_i - \theta_i)/\theta_i\big] \\ &= E\big\{\mathrm{Var}\big[(\tilde{\theta}_i - \theta_i)/\theta_i|\theta_i\big]\big\} + \mathrm{Var}\big\{E\big[(\tilde{\theta}_i - \theta_i)/\theta_i|\theta_i\big]\big\} \\ &= \big\{n_i \exp\big(2\mu + \sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big)^2 \cdot E(1/\theta_i + 1/\kappa) \\ &\quad + \exp\big(2\mu + \sigma^2\big)\big[\exp\big(\mu + 3\sigma^2/2\big)/\kappa + 1\big]^2 \cdot \mathrm{Var}(1/\theta_i)\big\} \\ &\quad \div \big[\exp\big(\mu + 3\sigma^2/2\big)/\kappa + 1 + n_i \exp\big(\mu + \sigma^2/2\big)\big(\exp\big(\sigma^2\big) - 1\big)\big]^2 \\ &= \big\{n_i \exp\big(2\mu + \sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big)^2\big[\exp\big(\sigma^2/2 - \mu\big) + 1/\kappa\big] \\ &\quad + \exp\big(2\sigma^2\big)\big(\exp\big(\sigma^2\big) - 1\big)\big[\exp\big(\mu + 3\sigma^2/2\big)/\kappa + 1\big]^2\big\} \\ &\quad \div \big[\exp\big(\mu + 3\sigma^2/2\big)/\kappa + 1 + n_i \exp\big(\mu + \sigma^2/2\big)\big(\exp\big(\sigma^2\big) - 1\big)\big]^2.\end{aligned}$$

APPENDIX C: PARAMETRIC BOOTSTRAP ALGORITHM TO APPROXIMATE $U_{i,\alpha}$

1. Generate an $r \times c$ spatial grid based on the heuristic optimal design.
2. Simulate insect counts from the sampled trees in the co-clusters from independent distributions of Negative Binomial$(\hat{\tilde{\theta}}_i, \hat{\kappa})$.
3. Fit the GLMM based on the simulated counts to obtain $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$.
4. With (μ, σ^2, κ) replaced by $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$ in (3.1) and (3.2), calculate $\hat{\tilde{\theta}}_i^* = \mathrm{eBLP}(\theta_i)$ and $\widehat{\mathrm{Var}}^*(\log\tilde{\theta}_i - \log\theta_i)$.
5. Calculate $U_i^* = (\log\hat{\tilde{\theta}}_i^* - \log\hat{\tilde{\theta}}_i)/\sqrt{\widehat{\mathrm{Var}}^*(\log\tilde{\theta}_i - \log\theta_i)}$.
6. Repeat Steps 2–5 $B = 1000$ times to obtain $U_i^{*(1)}, U_i^{*(2)}, \ldots, U_i^{*(B)}$.
7. Approximate $U_{i,\alpha}$ by the $100(1-\alpha)^{th}$ percentile of $U_i^{*(1)}, U_i^{*(2)}, \ldots, U_i^{*(B)}$.

[Received June 2011. Accepted March 2012. Published Online April 2012.]

REFERENCES

Barchia, I. M., Herron, G. A., and Gilmour, A. R. (2003), "Use of a Generalized Linear Mixed Model to Reduce Excessive Heterogeneity in Petroleum Spray Oil Bioassy Data," *Journal of Economic Entomology*, 96 (3), 983–989.

Bennett, K. E., Hopper, J. E., Stuart, M. A., West, M., and Drolet, B. S. (2008), "Blood-Feeding Behavior of Vesicular Stomatitis Virus Infected *Culicoides Sonorensis* (Diptera: Ceratopogonidae)," *Journal of Medical Entomology*, 45 (5), 921–926.

Bianchi, F. J. J. A., Goedhart, P. W., and Baveco, J. M. (2008), "Enhanced Pest Control in Cabbage Crops Near Forest in The Netherlands," *Landscape Ecology*, 23 (5), 595–602.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88 (421), 9–25.

Busygin, S., Prokopyev, O., and Pardalos, P. M. (2008), "Biclustering in Data Mining," *Computers & Operations Research*, 35 (9), 2964–2987.

Candy, S. G. (2000), "The Application of Generalized Linear Mixed Models to Multi-level Sampling for Insect Population Monitoring," *Environmental and Ecological Statistics*, 7 (3), 217–238.

Elias, S. P., Lubelczyk, C. B., Rand, P. W., Lacombe, E. H., Holman, M. S., and Smith, R. P. (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," *Journal of Medical Entomology*, 43 (6), 1142–1152.

Elston, D. A., Moss, R., Boulinier, T., Arrowsmith, C., and Lambin, X. (2001), "Analysis of Aggregation, a Worked Example: Numbers of Ticks on Red Grouse Chicks," *Parasitology*, 122 (5), 563–569.

Gotway, C. A., and Stroup, W. W. (1997), "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 2 (2), 157–178.

Gozé, E., Nibouche, S., and Deguine, J.-P. (2003), "Spatial and Probability Distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in Cotton: Systematic Sampling, Exact Confidence Intervals and Sequential Test," *Environmental Entomology*, 32 (5), 1203–1210.

Hartigan, J. A. (1972), "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, 67 (337), 123–129.

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 Savopoulou-Soultani, M. (2006), "Use of Geostatistical Analysis to Characterize the Spatial Distribution of *Lobesia botrana* (Lepidoptera: Tortricidae) Larvae in Northern Greece," *Environmental Entomology*, 35 (2), 497–506.

Madeira, S. C., and Oliveira, A. L. (2004), "Biclustering Algorithms for Biological Data Analysis: A Survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1 (1), 24–45.

Maoz, Y., Gal, S., Zilberstein, M., Izhar, Y., Alchanatis, V., Coll, M., and Palevsky, E. (2011), "Determining an Economic Injury Level for the Persea Mite *Oligonychus perseae*, a New Pest of Avocado in Israel," *Entomologia Experimentalis et Applicata*, 138 (2), 110–116.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.), Hoboken: Wiley.

Mechelen, I. V., Bock, H.-H., and Boeck, P. D. (2004), "Two-mode Clustering Methods: A Structured Overview," *Statistical Methods in Medical Research*, 13, 363–394.

Olejnik, S., Li, J., Supattathum, S., and Huberty, C. J. (1997), "Multiple Testing and Statistical Power With Modified Bonferroni Procedures," *Journal of Educational and Behavioral Statistics*, 22 (4), 389–406.

Paterson, S., and Lello, J. (2003), "Mixed Models: Getting the Best Use of Parasitological Data," *Trends in Parasitology*, 19 (8), 370–375.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006), "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data," *Bioinformatics*, 22 (9), 1122–1129.

Ramírez-Dávila, J. F., and Porcayo-Camargo, E. (2008), "Spatial Distribution of the Nymphs of *Jacobiasca Lybica* (Hemiptera: Cicadellidae) in a Vineyard in Andalucia, Spain," *Revista Colombiana de Entomologia*, 34 (2), 169–175.

Schotzko, D. J., and O'Keeffe, L. E. (1989), "Geostatistical Description of the Spatial Distribution of *Lygus hesperus* (Heteroptera: Miridae) in Lentils," *Journal of Economic Entomology*, 82 (5), 1277–1288.

Shah, P. (2006), "Sequential Sampling Methods Using Generalized Linear Models With Applications to Pest Density Estimation," Ph.D. Dissertation, University of California, Riverside, CA.

Takakura, K.-I. (2009), "Reconsiderations on Evaluating Methodology of Repellent Effects: Validation of Indices and Statistical Analyses," *Journal of Economic Entomology*, 102 (5), 1977–1984.

Williams, L., Schotzko, D. J., and McCaffrey, J. P. (1992), "Geostatistical Description of the Spatial Distribution of *Limonius Californicus* (Coleoptera: Elateridae) Wireworms in the Northwestern United States, With Comments on Sampling," *Environmental Entomology*, 21 (5), 983–995.

Zhang, Z. (2011), "Clustering: Algorithm, Optimization and Inference," Ph.D. Dissertation, University of California, Riverside, CA.