MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **15**, Number **4**, August **2018**

doi:10.3934/mbe.2018044

pp. **993–1010**

OPTIMAL DESIGN FOR DYNAMICAL MODELING OF PEST POPULATIONS

H. T. Banks, R. A. Everett, Neha Murad and R. D. White

Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695-8212, USA

J. E. Banks

Undergraduate Research Opportunities Center (UROC)
California State University, Monterey Bay
Seaside, CA 93955, USA

Bodil N. Cass and Jay A. Rosenheim

Department of Entomology and Nematology, Center for Population Biology
University of California, Davis
CA 95616, USA

(Communicated by Jia Li)

Abstract. We apply *SE*-optimal design methodology to investigate optimal data collection procedures as a first step in investigating information content in ecoinformatics data sets. To illustrate ideas we use a simple phenomenological citrus red mite population model for pest dynamics. First the optimal sampling distributions for a varying number of data points are determined. We then analyze these optimal distributions by comparing the standard errors of parameter estimates corresponding to each distribution. This allows us to investigate how many data are required to have confidence in model parameter estimates in order to employ dynamical modeling to infer population dynamics. Our results suggest that a field researcher should collect at least 12 data points at the optimal times. Data collected according to this procedure along with dynamical modeling will allow us to estimate population dynamics from presence/absence-based data sets through the development of a scaling relationship. These Likert-type data sets are commonly collected by agricultural pest management consultants and are increasingly being used in ecoinformatics studies. By applying mathematical modeling with the relationship scale from the new data, we can then explore important integrated pest management questions using past and future presence/absence data sets.

1. **Introduction.** Integrated pest management (IPM) is an ecosystem-based process for managing pests that interfere with or damage crops. It investigates long-term prevention of pests using a variety of methods, such as biological and cultural controls [15]. IPM research relies heavily on information gained from data sets, and recently, some entomologists have advocated the supplemental use of

2010 *Mathematics Subject Classification.* 34L30, 90C30, 34A55, 65L09.

Key words and phrases. Dynamic modeling, pest management, optimal experimental design, constrained optimization, ecoinformatics.

"ecoinformatics". Ecoinformatics studies address ecological questions using observational, preexisting data rather than experimental, researcher-generated data and often combine data sets from several sources into a larger data set [18, 19]. These sources can include farmers, pest management consultants (PMC), federal and state repositories, among others [18].

There are several weaknesses in experimental approaches that can be complemented by ecoinformatics. For example, due to cost limitations, experiments are often on a smaller scale, both spatially and temporally, while ecoinformatics approaches often reflect the scale of the farming that is being studied. The goals of IPM include improving crop yield for farmers; however, information drawn from experiments may only be relevant to a limited range of farming conditions. Ecoinformatics can include farmer participation from the start, and the farmers may be more confident in the recommendations that are generated from analyzing their own data [18, 19]. Although there are several benefits of applying ecoinformatics methods to IPM research, an important potential weakness to address is the *information content of the data*, which affects the accuracy of the resulting conclusions. Ecoinformatics data sets are often heterogeneous due to the variety of sources and sampling methods. In addition, pest densities and other variables of interest can be measured qualitatively rather than quantitatively (e.g., "trace", "low", "moderate", and "high" densities in Likert-type [16] data sets as opposed to population counts). There is, of course, a trade-off; collecting qualitative data is much more time efficient but can significantly reduce the information content in the data. For further discussions on the strengths and weaknesses of experimental and ecoinformatics data sets, and for a review on ecoinformatics in the context of agricultural entomology, see [18, 19, 12] and the references therein.

Our own efforts in dealing with such Likert type data sets arose in dealing with qualitative data sets such as those in [17]. In order to use mathematical models to detect trends in this type of data, *population counts* as well as corresponding Likert-type data are needed to establish a scale between these two types of data. This scale can then be applied to existing qualitative data sets. The optimal design for quantitative data collection (sampling strategies including how often and how much data to be collected) maximizes the accuracy in estimating population dynamics and is the major focus of this brief note.

The ability to perform this data conversion is an important step for determining the information content in farmer-generated ecoinformatics data sets. There are numerous methods to investigate quality (information content) of a data set, among them the use of dynamical models and related sensitivity as well as statistical uncertainty quantification tools. In this context, information content of a data set refers to the quality of the data with respect to *accurate estimation* of model parameters with an *acceptable statistical confidence* associated with these parameters. The parameters are first determined by solving an inverse problem such that the model solution best fits the data. A data set with high quality contains sufficient information to produce statistically accurate (such as acceptable confidence intervals or some other associated measure of uncertainty) parameter estimates. With accurate parameter estimates, a model solution can then realistically capture population trends, which can help, for example, investigate the minimum pesticide amount needed to reduce pest populations below an economic threshold. Examples of previous works using dynamical models to investigate information content in ecological data include [1, 2, 3, 7].

To illustrate the assessment of the quality of large ecoinformatics data sets, here we consider a subset of the data from [17] from a single PMC who collected repeated measures of citrus red mite (CRM), *Panonychus citri*, densities over multiple time points (e.g., longitudinal data – a necessity for applying dynamical systems to data). CRMs are citrus pests that extract cell sap from leaves and fruit, which causes yield loss and stippling that can reduce the grade of the fruit [13]. CRM populations gradually increase over the spring and then sharply decline during the hot summer months [11]. We wish to capture this growth trend using dynamical modeling, as this will allow us to evaluate the information content in the data. Investigating CRMs is a research pest management priority, specifically with respect to secondary outbreaks and the relationship between pest densities and loss of fruit quality/quantity.

The PMC generated data did not contain quantitative pest counts. Specifically, the subset considered here only provided CRM infestation proportion, defined as the proportion of leaves sampled that contain at least one CRM. That is,

$$\text{infestation proportion} = \frac{\text{infestation finding}}{\text{infestation sample}},$$

where infestation sample is the number of sample units (leaves) checked, and infestation finding is the number of sample units infested with one or more CRMs. This sampling method provides no quantitative information as to how many CRMs are present on each infested leaf. Thus, an increasing infestation proportion over the spring months provides only indirect information as to the dynamics of the CRM population. Therefore, we are not able to use only infestation proportion to analyze the seasonal trend in the data.

Previous work reports relationships between infestation proportion and a total population, which potentially could be used to convert our infestation proportion data to population counts. The authors in [14] develop a sampling plan to predict the total CRM population from the proportion of leaves infested with at least one adult female on the lower surface of a leaf. However, this relationship was developed based on lemon plants in Riverside and Ventura Counties in California. Thus this relationship may not be applicable to our data collected on oranges and mandarins in the San Joaquin Valley.

Therefore, we aim to apply optimal design methodology to determine when and how often to collect count data from similar fields in order to develop a relationship between CRM count and infestation proportion data, similar to that in [14]. That is, in this paper we aim to answer the following questions

1. For a set time period and a fixed number of data points, *when* should data be collected?
2. With optimized data collection time points, *how many* data points are needed?

Once data are collected according to the optimal design formulation, we can determine a relationship between CRM infestation proportion and total population. With this relationship, we can convert the infestation proportion data to population counts and apply our dynamical model to investigate the quality of the ecoinformatics data set as well as examine other pertinent IPM questions.

In Section 2 we introduce a simple CRM population model (primarily to illustrate ideas since a more sophisticated *validated* population model is not available) as well as the statistical model used in our optimal design formulation. The framework for this *SE*-optimal design is then given in Section 3, with the implementation of the

constrained optimization given in Section 4. Section 5 discusses computing standard errors (SE) using asymptotic theory for Monte Carlo simulations. The results are presented in Section 6 and conclusions are discussed in Section 7.

2. **Dynamical modeling of CRM populations.** Mathematical models are used to represent biological systems and investigate hypotheses regarding the biological processes. While a *mechanistic model* hypothesizes the relationships between different biologically interpretable variables and parameters, a *phenomenological model* solely aims to capture qualitative trends in the desired dynamics. We present a simple phenomenological CRM population model since here, we only aim to apply the model to optimal design methodology rather than hypothesize specific mechanisms of population growth and death. That is, we use a model that represents the general seasonal *trends* as represented in seasonal curves [11] and hence the model is *not* based on specific growth/death mechanisms from a *previously developed and validated* model. The simple mathematical model we use for this is given by

$$\frac{dx}{dt} = g(t)x\left(1 - \frac{x}{K}\right) - d(t)x, \tag{1a}$$

$$g(t) = a\sin(bt) \tag{1b}$$

$$d(t) = -c\cos(3bt) + c, \tag{1c}$$

$$x_0 = 100, \tag{1d}$$

with scalar observation process

$$f(t, \boldsymbol{\theta}) = x(t, \boldsymbol{\theta}), \tag{2}$$

with parameters $\boldsymbol{\theta} = (a, b, c, K) \in \mathbb{R}^p$, $f \in \mathbb{R}^m = \mathbb{R}$, and where x represents the number of CRMs. The CRM population is assumed to grow logistically with time-dependent intrinsic growth rate $g(t)$ and carrying capacity K. The CRM population death rate is also assumed to be time-dependent, given by $d(t)$. The tuning parameters, a, b, and c, adjust the shape of the intrinsic growth and death rate curves in this phenomenological model and hence do not have specific mechanistic-based meaning. The simple functions $g(t)$ and $d(t)$ were chosen so that the CRM dynamics in a 7 month season (January - July) generally reflect those reported in biological literature [11, 12] with a minimal number of parameters. Other simple functions commonly used in modeling, such as polynomials, can depend on a larger number of parameters, which generally require more data to estimate. The model solution for nominal parameters $\boldsymbol{\theta_0} = (a_0, b_0, c_0, K_0) = (0.12, 0.015, 0.025, 250)$ is given in Figure 2b and represents what studies suggest might be the dynamics of a typical infestation period [11, 12].

In order to account for the uncertainty we would expect in observational data, we consider the following statistical error model

$$Y(t) = f(t, \boldsymbol{\theta}_0) + \mathcal{E}(t), \tag{3}$$

where $Y(t)$ is a random variable, $\boldsymbol{\theta}_0$ is the nominal parameter vector, and $\mathcal{E}$ is assumed to be independent and identically distributed with mean 0 and variance σ_0^2. A realization of the statistical error model is given by

$$y(t) = f(t, \boldsymbol{\theta}_0) + \epsilon(t), \quad t \in [0, T], \tag{4}$$

where ϵ is a specific realization of the random variable $\mathcal{E}$.

3. ***SE* optimal design formulation.** We aim to determine the sampling times of experiments in order to maximize the information content in the data collected. In order to explain the optimal design methodology, we begin by giving an intuitive explanation of information content (Subsection 3.1). With this, we then provide the motivation for the specific type of optimal design implemented here (Subsection 3.2).

3.1. **Information content.** In this context, information content refers to the quality of the data in regards to estimating model parameters. That is, data with high information content allow us to accurately estimate parameters as well as attach high degrees of statistical confidence to these parameters. With this, one can hope to infer valuable information about the actual population trends.

We first discuss the motivation behind the SE-optimal design formulation. That is, we discuss how areas of high information content are determined. For intuition, let us consider the effect that the parameters have on the model solution (i.e., sensitivity of the model solution with respect to the parameters over time). Figures 1a - 1d depict these sensitivities. From this, one can see that the sensitivity of the model solution to a given parameter varies over time. Data taken at times where the solution is more sensitive to a given parameter, correspond to more accurate estimation of that parameter. For instance, consider the sensitivity of the carrying capacity, K, given in Figure (1d). One can see that as the season progresses, the sensitivity of the solution with respect to K steadily increases until it reaches its maximum towards the end of the season. This makes sense as one expects to have less information about the carrying capacity of the population early on, but attain the most information about the carrying capacity when the population reaches its maximum (around day 150). After this peak, there is little additional information gained, which corresponds to the decrease in sensitivity observed at these later times.

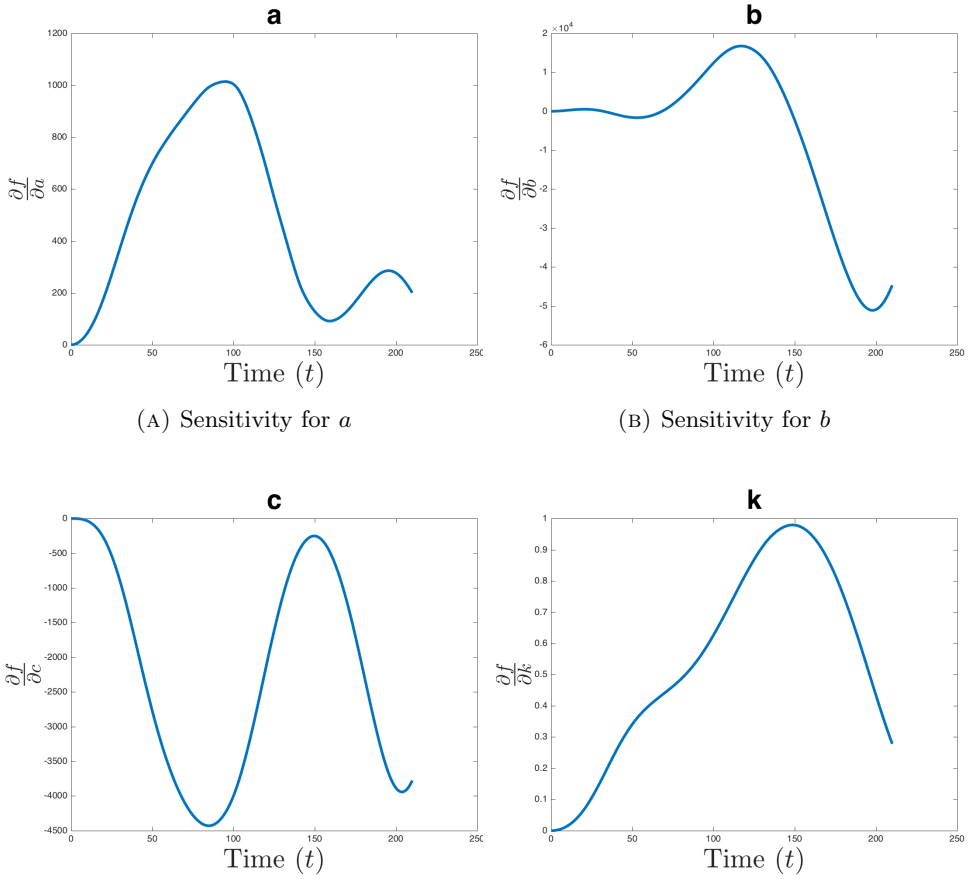
Among possible optimal design formulations (D-optimal, E-optimal, SE-optimal, etc. [5, 6, 8]), it is common to base the design criterion on the Fisher Information Matrix (FIM), as this indirectly includes information about the sensitivities. Section 3.2 discusses the FIM as well as the specific criterion for the SE-optimal design formulation used here.

3.2. **Optimal design criterion.** Although sensitivities play a role in determining information content, the individual sensitivities do not solely determine the optimal sampling times. Rather, the criterion takes into account a combination of the effects of sensitivities through the FIM. A derivation following [5, 6] is given next that explains how minimizing a criterion dependent on the FIM determines the optimal sampling times.

Given data corresponding to a distribution of sampling times, $P(t)$, one often evaluates how accurately a model solution fits these data via a weighted least squares cost functional. For instance, consider the error functional given below

$$J(y, \boldsymbol{\theta}) = \int_0^T \frac{1}{\sigma^2(t)} (y(t) - f(t, \boldsymbol{\theta}))^2 dP(t). \tag{5}$$

Note that this error functional represents a more general case where the variance in the data can change over time (although in our problem variance is assumed constant). The lower the value of J, the more closely the model fits the data. Our question is *what distribution of sampling times can produce the smallest J value*?



(A) Sensitivity for a

(B) Sensitivity for b

(C) Sensitivity for c

(D) Sensitivity for K

FIGURE 1. Traditional sensitivities for model parameters.

Since the goal is to determine the optimal sampling times *prior to data collection*, we wish to use (5) to develop a minimization criterion that is based on the mathematical model and is independent of the data. Recall, the statistical model is of the form

$$y(t) = f(t, \boldsymbol{\theta}_0) + \epsilon(t). \tag{6}$$

Then expanding $f(t, \boldsymbol{\theta})$ about the nominal parameter set $\boldsymbol{\theta}_0$ using a Taylor Series, we obtain

$$f(t, \boldsymbol{\theta}) \approx f(t, \boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} f(t, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{7}$$

where $\nabla_{\boldsymbol{\theta}}$ is given by $[\partial_{\boldsymbol{\theta}_1}, \ldots, \partial_{\boldsymbol{\theta}_p}]$. Note that $\nabla_{\boldsymbol{\theta}} f$ is an $1 \times p$ matrix, which gives the sensitivity of the solution with respect to the parameters. Now let us substitute (6) and (7) into the functional given in (5), resulting in the modified functional

$$\tilde{J}(y, \boldsymbol{\theta}) = \int_0^T \frac{1}{\sigma^2(t)} \Big(\epsilon(t) - \nabla_{\boldsymbol{\theta}} f(t, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \Big)^2 dP(t), \tag{8}$$

where $J \approx \tilde{J}$ in a neighborhood of $\boldsymbol{\theta}_0$. Note that

$$\nabla_{\boldsymbol{\theta}}\tilde{J}(y,\boldsymbol{\theta}) = -2\int_0^T \frac{1}{\sigma^2(t)}\Big(\epsilon(t) - \nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\Big)\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)dP(t).$$

We observe that a minimum argument $\tilde{\boldsymbol{\theta}}$ in cost functional (8) (tacitly assumed to occur in the interior of the set of possible of values) implies that

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}}\tilde{J}(y,\tilde{\boldsymbol{\theta}})\\
&= -2\int_0^T \frac{1}{\sigma^2(t)}\Big(\epsilon(t) - \nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\Big)\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)dP(t)\\
&= -2\int_0^T \frac{1}{\sigma^2(t)}\Big(\epsilon(t)\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0) - (\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)\Big)dP(t) = \mathbf{0}_{1\times p},
\end{aligned} \tag{9}$$

or equivalently

$$\int_0^T \frac{\epsilon(t)}{\sigma^2(t)}\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)dP(t) - (\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\int_0^T \frac{1}{\sigma^2(t)}\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(t,\boldsymbol{\theta}_0)dP(t) = \mathbf{0}_{1\times p}. \tag{10}$$

We see that this equation contains the Generalized Fisher Information Matrix (GFIM), defined by

$$F(P,\boldsymbol{\theta}_0) = \int_0^T \frac{1}{\sigma_0^2(s)}\nabla_{\boldsymbol{\theta}} f(s,\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(s,\boldsymbol{\theta}_0)dP(s). \tag{11}$$

Since our optimal mesh is considered to be a discrete set of time points, we can now introduce a discretization of the sampling distribution $P(t)$. Without loss of generality we can consider these distributions as probability measures on $[0,T]$, where the set of all such measures is denoted $P(0,T)$. Suppose for points $\tau = \{t_i\}_{i=1}^N \in [0,T]$ we have

$$P_\tau = \sum_{i=1}^N \Delta_{t_i}, \tag{12}$$

where Δ_{t_i} is the Heaviside function (with the derivative being the Dirac delta function) with atom at $\{t_i\}$ (see Appendix 7). That is,

$$\Delta_{t_i}(t) = \begin{cases} 1, & t \geq t_i \\ 0, & t < t_i. \end{cases} \tag{13}$$

Considering the measure P given above, we have the discrete version of (10) given by

$$\sum_{i=1}^N \frac{\epsilon(t_i)}{\sigma^2(t_i)}\nabla_{\boldsymbol{\theta}} f(t_i,\boldsymbol{\theta}_0) - (\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\sum_{i=1}^N \frac{1}{\sigma^2(t_i)}\nabla_{\boldsymbol{\theta}} f(t_i,\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(t_i,\boldsymbol{\theta}_0) = \mathbf{0}_{1\times p}. \tag{14}$$

We observe that this contains the discrete form Fisher Information Matrix given by

$$F(P_\tau,\boldsymbol{\theta}_0) = \sum_{i=1}^N \frac{1}{\sigma^2(t_i)}\nabla_{\boldsymbol{\theta}} f(t_i,\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(t_i,\boldsymbol{\theta}_0), \tag{15}$$

which is tacitly assumed to be of full rank. Now consider that we want $\tilde{\boldsymbol{\theta}}$ to be as similar to $\boldsymbol{\theta}_0$ as possible and solve for $(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T$ in (14):

$$(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T = bF^{-1}, \quad \text{or} \quad (\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = F^{-1}b^T, \tag{16}$$

where $b = b(P_\tau, \boldsymbol{\theta}_0) = \sum_{i=1}^{N} \frac{\epsilon(t_i)}{\sigma^2(t_i)} \nabla_{\boldsymbol{\theta}} f(t_i, \boldsymbol{\theta}_0)$. We see that b contains the observational random error term, ϵ, on which we would not want to base our design.

From (16) one can see why a minimization criterion for the optimal design formulation is based on F^{-1}. Now let us recall the optimal design problem. That is, we wish to determine the optimal $\hat{P}_\tau$ such that, for $\mathcal{J} : \mathbb{R}^{p\times p} \to \mathbb{R}^+$,

$$\mathcal{J}(F(\hat{P}_\tau, \boldsymbol{\theta}_0)) = \min_{P_\tau \in P(0,T)} \mathcal{J}(F(P_\tau, \boldsymbol{\theta}_0)). \tag{17}$$

Specifically for *SE*-optimal design, $\mathcal{J}_{SE}$ is given by

$$\mathcal{J}_{SE}(F) = \sum_{k=1}^{p} \frac{1}{\boldsymbol{\theta}_{0,k}^2} (F^{-1})_{kk}. \tag{18}$$

Minimizing this cost functional corresponds to minimizing the sum of the squared normalized standard errors, where standard errors are used to calculate confidence intervals for parameter estimates (see Section 5.1).

4. **Constrained optimization and *SE* design implementation.** The *SE*-optimal design computational method utilizes a constrained optimization to determine the mesh of time points, $\tau^* = \{t_i^*\}, i = 1, \ldots, N$, that satisfy

$$\mathcal{J}(F(P_{\tau^*}, \boldsymbol{\theta}_0)) = \min_{\tau \in \mathcal{T}} \mathcal{J}(F(P_\tau, \boldsymbol{\theta}_0)), \tag{19}$$

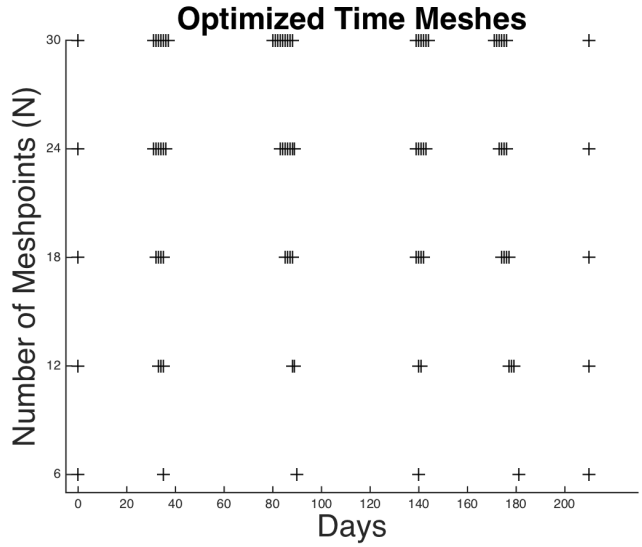
where $\mathcal{T}$ is the set of all time meshes such that $0 < t_2 < \cdots < t_{N-1} < T$. The algorithm used to implement this constrained optimization is MATLAB's *fmincon*. Since the optimal mesh should contain 0 and T, which are assumed to be known, we optimize over $N-2$ parameters. To enforce the linear time mesh constraint in *fmincon*, we use the following linear system

$$A\mathbf{t} \le \mathbf{b}, \tag{20}$$

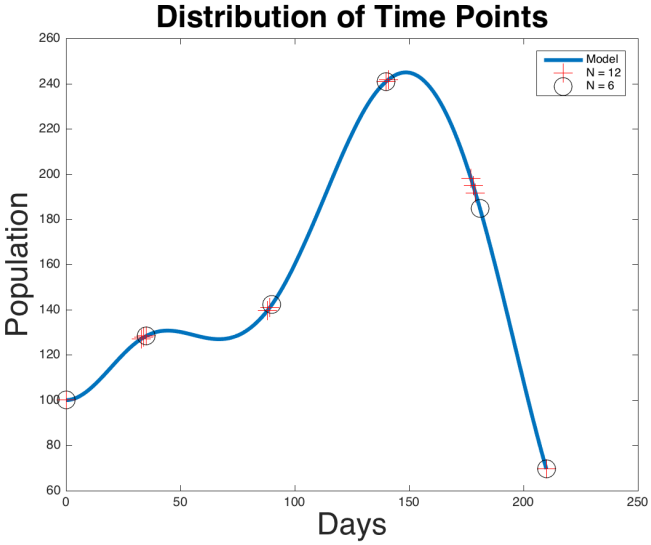
where A is an $(N-1)\times(N-2)$ matrix, $\mathbf{t}$ is an $(N-2)\times 1$ time vector, and $\mathbf{b}$ is an $(N-1)\times 1$ vector. For this implementation, (20) has the following form:

$$\begin{bmatrix} -1 & 0 & & \\ 1 & -1 & \ddots & \\ 0 & \ddots & \ddots & 0 \\ & \ddots & 1 & -1 \\ & & 0 & 1 \end{bmatrix} \begin{bmatrix} t_2 \\ \\ \vdots \\ \\ t_{N-1} \end{bmatrix} \le \begin{bmatrix} -1 \\ -1 \\ \\ \vdots \\ \\ T-1 \end{bmatrix}. \tag{21}$$

This constraint forces the first optimized mesh point to be greater than or equal to 1, the final optimized mesh point to be less than or equal to $T-1$, and all interior mesh points to be at least one day apart (since this is reasonable in the field). Furthermore, note that although we are dealing with discrete days, we do not force this in the optimization. Once the optimal mesh is determined, we round to the nearest whole number. This seems reasonable in practice since we are not concerned with what time of day sampling occurs. For this experiment, we consider grids with $N = 6$, 12, 18, 24, and 30. This corresponds to sampling 6 times in the sampling season (January through July) and considers how doubling the number of samples improves our ability to estimate parameters accurately. Since uniform sampling may be more feasible in practice, we compare the standard errors corresponding to the optimal grids to those of uniform grids. Figure 2a depicts the distribution



(A) Optimized meshes for N mesh points

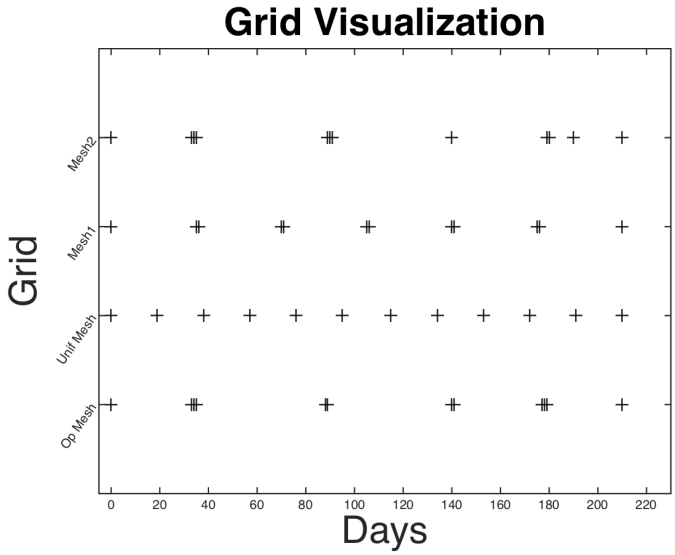


(B) Sample meshes on model solution

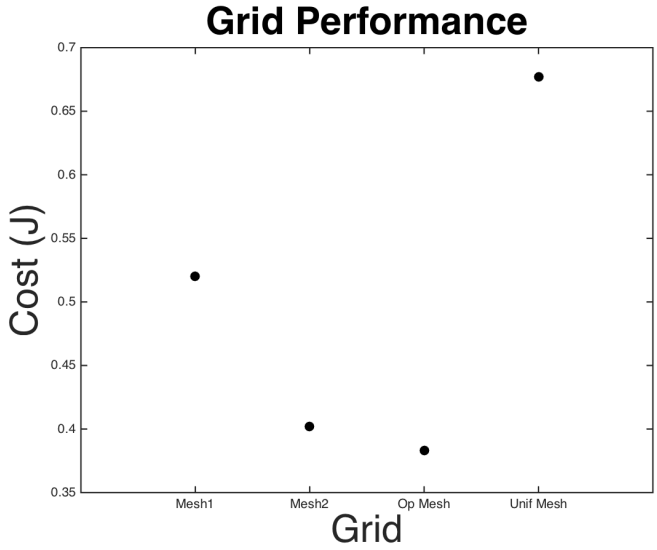
FIGURE 2. Optimized meshes resulting from *SE*-optimal implementation.

of sampling times for the optimized grids. Figure 2b shows these distributions for $N = 6$ and $N = 12$ along the solution curve.

We note that the optimized time meshes cluster to areas of high information content, based on the cost functional in (18). To provide intuition as to why this occurs, we plot in Figure 3a the cost value (J) for the optimal mesh, the uniform mesh, and two hypothetical meshes for $N = 12$. The hypothetical meshes were designed to have clustering similar but not identical to the optimal mesh. As expected, the optimized mesh produces the smallest cost value, while Mesh 2 (the most similar to



(A) Grids with $N = 12$ mesh points



(B) Performance for Meshes

FIGURE 3. Relationship between sampling distribution and corresponding performance (cost).

the optimal mesh) has the second lowest cost. We observe that the uniform mesh has the largest cost. In the context of inverse problems, it is advantageous to have multiple samples in time periods with high information content.

5. **Standard error methodology.** We first implement the constrained optimization scheme using the *SE* design formulation to determine the optimal distribution

of sampling points $\tau^* = \{t_j^*\}_{j=1}^N$ for fixed values of N. We then generate simulated data corresponding to these optimal meshes as well as to uniform meshes and compare standard errors for different sampling distributions. The following section describes the method for computing asymptotic standard errors for scalar models such as the one given in model (1).

5.1. **Asymptotic theory for computing standard errors.** Consistent with the statistical error model given in equation 4, we estimate our parameters by solving an inverse problem with an ordinary least squares (OLS) formulation, following [9, 10]. The OLS estimator is given by

$$\boldsymbol{\Theta}_{OLS} = \boldsymbol{\Theta}_{OLS}^N = \text{argmin}_{\boldsymbol{\theta}} \sum_{j=1}^{N} [Y_j - f(t_j, \boldsymbol{\theta})]^2, \tag{22}$$

which is estimated as

$$\hat{\boldsymbol{\theta}}_{OLS} = \hat{\boldsymbol{\theta}}_{OLS}^N = \text{argmin}_{\boldsymbol{\theta}} \sum_{j=1}^{N} [y_j - f(t_j, \boldsymbol{\theta})]^2. \tag{23}$$

Since the dependence of our estimate on the OLS formulation is understood, the OLS subscript notation will be dropped. Next, we compute the sensitivity matrix

$$\chi_{j,k} = \frac{\partial f(t_j, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_k}, \quad j = 1, \ldots, N, \quad k = 1, \ldots, p, \tag{24}$$

which is done using the complex step method [4]. That is,

$$\frac{\partial f(t_j, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_k} = \frac{\partial x(t_j, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_k} = \frac{\text{Im}(x(t_j, \hat{\boldsymbol{\theta}} + ih\boldsymbol{e}_k))}{h}, \tag{25}$$

where h is size of the perturbation, $\boldsymbol{e}_k$ is the k^{th} unit vector in $\mathbb{R}^p$, and i is the imaginary unit. Note that $\chi = \chi^N$ is an $N \times p$ matrix. The true, constant variance is given by

$$\sigma_0^2 = \frac{1}{N} E\left[\sum_{j=1}^{N} [Y_j - f(t_j, \boldsymbol{\theta}_0)]^2\right]. \tag{26}$$

We can estimate this variance by

$$\hat{\sigma}^2 = \frac{1}{N-p}\left[\sum_{j=1}^{N} [y_j - f(t_j, \hat{\boldsymbol{\theta}})]^2\right]. \tag{27}$$

The true covariance matrix is approximately given by

$$\Sigma_0^N \approx \sigma_0^2 \left[\chi^T(\boldsymbol{\theta}_0)\chi(\boldsymbol{\theta}_0)\right]^{-1}, \tag{28}$$

and the true Fisher Information Matrix (FIM) is given by

$$F = F(\tau, \boldsymbol{\theta}_0) = (\Sigma_0^N)^{-1}. \tag{29}$$

When $\boldsymbol{\theta}_0$ and σ_0^2 are unknown, the covariance matrix is estimated by

$$\hat{\Sigma}^N(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 \left[\chi^T(\hat{\boldsymbol{\theta}})\chi(\hat{\boldsymbol{\theta}})\right]^{-1}, \tag{30}$$

for which the corresponding estimate of the FIM is

$$\hat{F} = F(\tau, \hat{\boldsymbol{\theta}}) = (\hat{\Sigma}^N(\hat{\boldsymbol{\theta}}))^{-1}. \tag{31}$$

Then, the asymptotic standard errors are given by

$$SE_k(\boldsymbol{\theta}_0) = \sqrt{(\Sigma_0^N)_{kk}}, \quad k = 1, \ldots, p, \tag{32}$$

which are estimated by

$$SE_k(\hat{\boldsymbol{\theta}}) = \sqrt{(\hat{\Sigma}^N(\hat{\boldsymbol{\theta}}))_{kk}}, \quad k = 1, \ldots, p. \tag{33}$$

The confidence interval for parameter estimate $\hat{\theta}_k$ with a confidence level of $100(1-\alpha)\%$, is given by

$$[\hat{\theta}_k - t_{1-\alpha/2} SE_k(\hat{\boldsymbol{\theta}}), \hat{\theta}_k + t_{1-\alpha/2} SE_k(\hat{\boldsymbol{\theta}})], \tag{34}$$

where $\alpha \in [0, 1]$ and $t_{1-\alpha/2}$ is computed from the Student's t distribution with $N-p$ degrees of freedom.

5.2. **Monte Carlo methods for asymptotic standard errors.** Monte Carlo (MC) trials can be used to examine the average asymptotic behavior of the standard errors. This accounts for the variability in residual errors in *simulated* data sets (as we have indicated earlier, no experimental *quantitative* data sets are available to test our results). For each Monte Carlo trial, data are simulated as

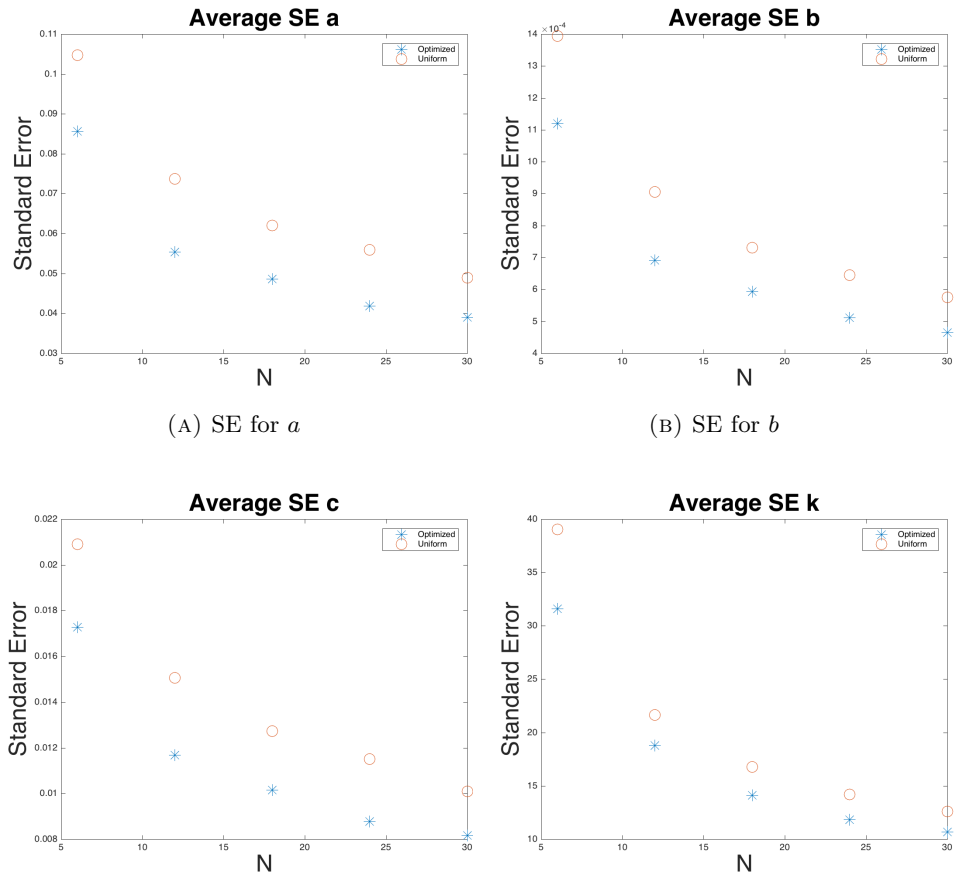
$$y_j = f(t_j, \boldsymbol{\theta}_0) + \epsilon_j, \quad j = 1, \ldots, N, \tag{35}$$

where $\boldsymbol{\theta}_0$ is the nominal parameter set, N corresponds to the number of time points in the optimal mesh $\{t_j^*\}_{j=1}^N$, and ϵ_j is a realization of $\mathcal{E}_j \sim \mathcal{N}(0, \sigma_0^2)$ for $\sigma_0 = 20$. For each trial, parameters are estimated and standard errors calculated using the OLS procedure described in Section 5.1. The average standard errors and parameter estimates are calculated over 1000 Monte Carlo trials. This provides the average performance of each optimal grid over 1000 noisy data sets.

6. **Results.** In Figure 4, the average standard errors are given for each parameter over 1000 Monte Carlo trials for both the optimized and uniform time meshes corresponding to $N = 6$, 12, 18, 24, and 30. Observe that for each N, the standard errors for the optimized grids are lower than those of the uniform grids, which is expected. Also note that as N increases, the standard errors for both the optimized and uniform grids decrease. It should be noted that although the optimized grids consistently perform better than the uniform grids, the standard errors for both might be considered acceptable, as they are all at least one order magnitude smaller than their corresponding parameter value.

In Figure 5, 95% confidence intervals are given using the average standard errors for each parameter corresponding to the optimal grids for $N = 6$, 12, 18, 24, and 30. The average parameter estimate is given by the dot at the center of each interval and is close to the nominal value. We observe that as N increases, the confidence intervals for each parameter become more narrow, with the most substantial decrease in interval width being between $N = 6$ and $N = 12$. This suggests that as the number of mesh points on the optimal grid increase, we are able to estimate the parameters with increasing accuracy.

From Figures 4 and 5 we see that data collected according to the optimal grid design provide acceptable standard errors, which allow us to be confident in the parameter estimates. However, we see that there is not a substantial improvement in standard errors and confidence intervals for $N > 12$. This decrease in improvement as N increases is reasonable due to the fact that sampling times cluster around areas of high information content, leading to a limiting effect in improvement.



(A) SE for a

(B) SE for b

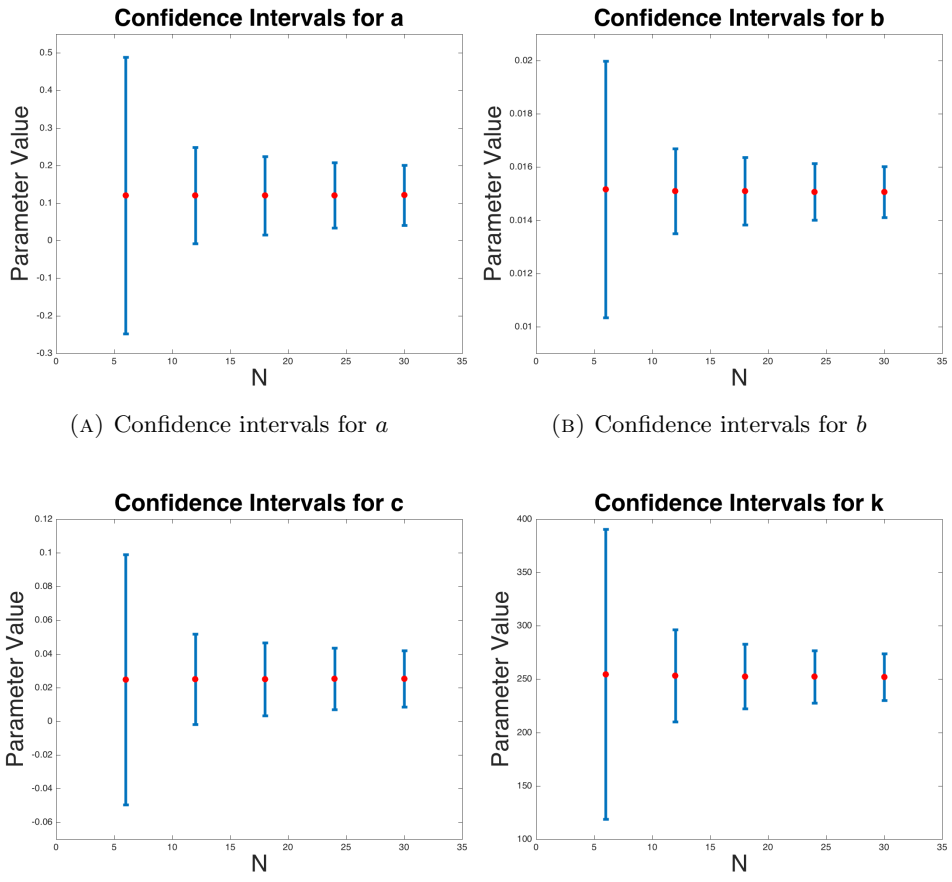
(C) SE for c

(D) SE for K

FIGURE 4. Average standard errors (over 1000 MC trials) for each parameter, comparing optimized versus uniform grids for N = 6, 12, 18, 24, and 30.

7. **Discussion.** We have determined an optimal design with regards to when observational CRM data should be collected. This optimal design criterion provides that data are collected in such a way that parameters can more confidently be estimated. Population count data collected according to the optimal grids would permit the use of dynamical modeling to infer CRM population sizes over a growing season. More importantly, with simultaneously collected corresponding proportional data (collected at the same time and with regards to the same sample unit), a *scaling relationship* between the population size in counts and corresponding proportional data could be estimated. This could allow us to make use of the current and future farmer-generated data sets consisting of only proportional data to develop and validate a suite of *mechanistic mathematical models* for use in investigating pest population dynamics using the broad ecoinformatics datasets.

We first addressed the question, *given a fixed number of data collection points, when are the optimal times to collect data*? To do this, we use the *SE*-optimal design framework for fixed $N = 6, 12, 18, 24$, and 30 to obtain the optimal sampling grid.



(A) Confidence intervals for a

(B) Confidence intervals for b

(C) Confidence intervals for c

(D) Confidence intervals for K

FIGURE 5. Confidence intervals for each parameter for N = 6, 12, 18, 24, and 30 on the optimized grids.

We observed that the optimal sampling time points tend to aggregate in areas of high information content, resulting in clustered time points. In addition, this clustering could be beneficial when collecting the data; a field researcher would only need to collect data for intermittent time periods compared to uniformly throughout the entire growing season.

The next question we considered is *given these optimal meshes, how much data should a field researcher collect*? We analyzed the performance of these meshes by comparing the standard errors of parameter estimates corresponding to each grid. The parameters were estimated using OLS methodology and MC simulations. As expected, a higher number of data points coincides with lower standard errors, with limiting improvement. In addition, the optimized grid performs better than uniform grids of the same size. We felt this was an important comparison as uniform sampling is often the procedure for research data collection in the field.

In order to further determine how much data are adequate for dynamical modeling, we calculated confidence intervals for the estimated parameters. It is clearly seen that there is *no significant decrease in confidence interval width for* $N > 12$.

Since there are reasonable standard errors for all N examined, dynamical modeling could be beneficial with as few as 6 data points at optimal times. However, we recommend a minimum of 12 data points due to the significant decreases in the size of confidence intervals between $N = 6$ and $N = 12$. The days at which these samples should be taken are given as [0 33 34 35 88 89 140 141 177 178 179 210], where day 0 corresponds to January 1st, and day 210 corresponds to approximately July 31st in the examples considered here.

Answering the optimal sampling distribution questions (when and how much data to collect) is dependent, of course, upon the mathematical model chosen to represent the population dynamics. For example, it might be expected that growth/death rates may depend on density of the pests and hence a corresponding model (even a phenomenological one such as (1)) would require density dependent coefficients. Also, our phenomenological model solution represents only typical dynamics observed in a single growing season. To account for more realistic, time-varying, biological factors such as weather, predator-prey interactions, etc., a more mechanistic model would need to be developed and validated. Thus, we *emphasize the importance of interdisciplinary collaboration to pursue all aspects of the efforts represented here.*

Being able to infer *population level dynamic information* from proportional data collected by farmers would allow us to investigate important questions relating to ecoinformatics. Presence/absence sampling is more time efficient compared to counting individuals, which enables the collection of a larger volume of data (both spatially and temporally). This facilitates more timely pest management decisions. Once a scaling relationship between count and proportional data is estimated, large proportional data sets in combination with mathematical modeling can be used to investigate problems such as the minimal number of pesticide treatments needed while not reducing crop yield. In addition, a better understanding of crop vulnerability to pest damage over time could help define a window of crop sensitivity in the growing season. Furthermore, we could investigate the impact of pests on mandarin varieties, which make up a rapidly growing part of citrus production in the San Joaquin Valley, CA. To date, there have been few formal investigations into this impact, making it a meaningful problem to pursue in interdisciplinary efforts.

Appendix. In Section 3.2 the notion of a cost functional dependent on a distribution, $P(t)$, is introduced in equation (5). We then discuss the Generalized Fisher Information Matrix, where a discretization of the distribution provides us with the discrete Fisher Information Matrix. In this section of the appendix we provide the mathematical details for using the discretization of a distribution (P_τ) via Heaviside functions to go from equation (11) to equation (15) (the GFIM to the FIM).

We begin by introducing the Heaviside function with atom at $\{t_i\}$ (Figure 6a) is defined as

$$\Delta_{t_i}(t) = \begin{cases} 1, & t \geq t_i \\ 0, & t < t_i, \end{cases} \tag{36}$$

with derivative given by the Dirac delta "function" (Figure 6b):

$$\frac{d}{dt}\Delta_{t_i}(t) = \delta_{t_i}(t),$$

where

$$\delta_{t_i}(t) = \begin{cases} +\infty, & t = t_i \\ 0, & t \neq t_i. \end{cases}$$

Properties of the Dirac delta function include

$$\int_{-\infty}^{\infty} \delta_{t_i}(s)ds = 1,$$

and

$$\int_{-\infty}^{\infty} f(t)\delta_{t_i}(t)dt = f(t_i).$$

Consider points $\tau = \{t_i\}_{i=1}^N \in [0, T]$, and define

$$P_\tau(t) = \sum_{i=1}^{N} \Delta_{t_i}(t), \tag{37}$$

which is plotted in Figure 6c.

The derivative of $P_\tau(t)$ (Figure 6d) is given by

$$P'_\tau(t) = \frac{d}{dt}P_\tau(t) = \frac{d}{dt}\left(\sum_{i=1}^{N}\Delta_{t_i}(t)\right) = \sum_{i=1}^{N}\left(\frac{d}{dt}\Delta_{t_i}(t)\right) = \sum_{i=1}^{N}\delta_{t_i}(t).$$

Consider the following for some function $f(t)$ and distribution of sampling times $P(t)$:

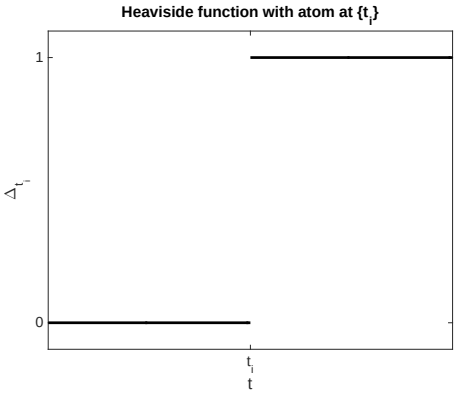
$$\int_0^T f(t)dP(t) = \int_0^T f(t)P'(t)dt.$$

Letting $P = P_\tau$, we have

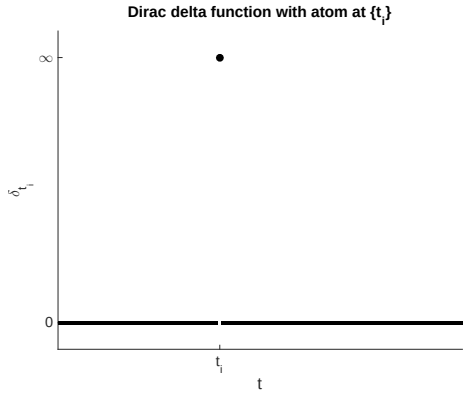
$$\begin{aligned} \int_0^T f(t)dP_\tau(t) &= \int_0^T f(t)P'_\tau(t)dt \\ &= \int_0^T f(t)[\delta_{t_1}(t) + \cdots + \delta_{t_N}(t)]dt \\ &= \int_0^T f(t)\delta_{t_1}(t)dt + \cdots + \int_0^T f(t)\delta_{t_N}(t)dt \\ &= \sum_{i=1}^{N} f(t_i). \end{aligned}$$

With this, one can see beginning with GFIM and introducing a distribution discretized as above we have the following

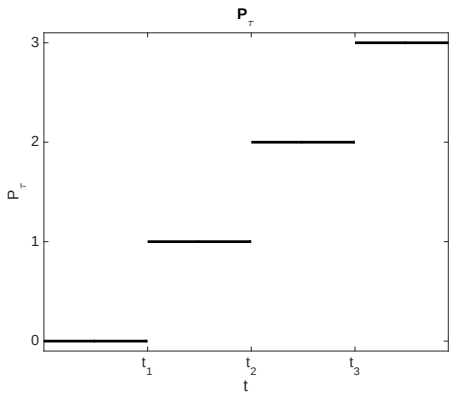
$$\begin{aligned} F(P, \boldsymbol{\theta}_0) &= \int_0^T \frac{1}{\sigma_0^2(s)}\nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)dP(s) \\ &= \int_0^T \frac{1}{\sigma_0^2(s)}\nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)P'(s)ds \\ \implies F(P_\tau, \boldsymbol{\theta}_0) &= \int_0^T \frac{1}{\sigma_0^2(s)}\nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)P'_\tau(s)ds \\ &= \int_0^T \frac{1}{\sigma_0^2(s)}\nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} f(s, \boldsymbol{\theta}_0)[\delta_{t_1}(s) + \cdots + \delta_{t_N}(s)]ds \\ &= \sum_{i=1}^{N} \frac{1}{\sigma^2(t_i)}\nabla_{\boldsymbol{\theta}} f(t_i, \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} f(t_i, \boldsymbol{\theta}_0). \end{aligned}$$



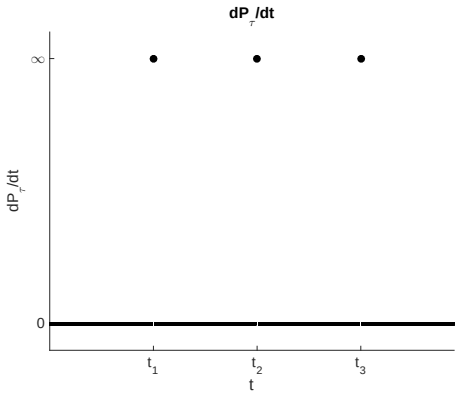
(A) Heaviside function with atom at $\{t_i\}$



(B) Dirac delta function with atom at $\{t_i\}$



(C) P_τ



(D) $\frac{dP_\tau}{dt}$

FIGURE 6. Heaviside functions and Dirac delta "functions".

Acknowledgments. This research was supported in part by the National Institute on Alcohol Abuse and Alcoholism under grant number 1R01AA022714-01A1, in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-15-1-0298, and in part by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) under grant number P200A120047.

REFERENCES

[1] H. T. Banks, J. E. Banks, R. A. Everett and J. D. Stark, An adaptive feedback methodology for determining information content in stable population studies, *Mathematical Biosciences and Engineering*, **13** (2016), 653–671.

[2] H. T. Banks, J. E. Banks, J. Rosenheim and K. Tillman, Modelling populations of *Lygus hesperus* cotton fields in the San Joaquin Valley of California: The importance of statistical and mathematical model choice, *Journal of Biological Dynamics*, **11** (2017), 25–39.

[3] H. T. Banks, J. E. Banks, N. Murad, J. A Rosenheim and K. Tillman, Modelling pesticide treatment effects on *Lygus hesperus* in cotton fields, *CRSC-TR15-09, Center for Research in Scientific Computation*, N. C. State University, Raleigh, NC, September, 2015; *Proceedings, 27 th IFIP TC7 Conference 2015 on System Modelling and Optimization,* L. Bociu et al (Eds.) CSMO 2015 IFIP AICT, **494** (2017), 1–12, Springer.

[4] H. T. Banks, K. Bekele-Maxwell, L. Bociu, M. Noorman and K. Tillman, The complex-step method for sensitivity analysis of non-smooth problems arising in biology, *Eurasian Journal of Mathematical and Computer Applications*, **3** (2015), 15–68.
[5] H. T. Banks, A. Cintron-Arias and F. Kappel, Parameter selection methods in inverse problem formulation, CRSC-TR10-03, N.C. State University, February, 2010, Revised, November, 2010; in *Mathematical Modeling and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems*, (J. J. Batzel, M. Bachar, and F. Kappel, eds.), 43–73, Lecture Notes in Mathematics, 2064, Springer-Verlag, Berlin 2013.
[6] H. T. Banks, S. Dediu, S. L. Ernstberger and F. Kappel, Generalized sensitivities and optimal experimental design, *Journal of Inverse and Ill-posed Problems*, **18** (2010), 25–83.
[7] H. T. Banks and M. L. Joyner, *Information Content in Data Sets: A Review of Methods for Interrogation and Model Comparison*, CRSC-TR17-14, N. C. State University, Raleigh, NC, June, 2017.
[8] H. T. Banks, K. Holm and F. Kappel, Comparison of optimal design methods in inverse problems, *Inverse Problems*, **27** (2011), 075002, 31 pp.
[9] H. T. Banks, S. Hu and W. C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, CRC Press, New York, 2014.
[10] H. T. Banks and H. T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, New York, 2009.
[11] C. C. Childers and T. R. Fasulo, Citrus red mite, *Gainesville: University of Florida Institute of Food and Agricultural Sciences*, ENY817, 1992. http://ufdc.ufl.edu/IR00004619/00001
[12] S. H. Dreistadt, *Review of Integrated Pest Management for Citrus,* 3rd ed, Journal of Agricultural & Food Information, University of California Division of Agriculture and Natural Resources, Publication 3303, 2012.
[13] L. Ferguson and E. E. Grafton-Cardwell, *Citrus Production Manual*, University of California Agriculture and Natural Resources, Publication 3539, 2014.
[14] V. P. Jones and M. P. Parrella, Intratree regression sampling plans for the citrus red mite (Acari: Tetranychidae) on lemons in southern California, *Journal of Economic Entomology*, **77** (1984), 810–813.
[15] M. Kogan, Integrated pest management: historical perspectives and contemporary developments, *Annual Review of Entomology*, **43** (1998), 243–270.
[16] R. Likert, A technique for the measurement of attitudes, *Archives of Psychology*, **22** (1932), p55.
[17] G. Livingston, L. Hack, K. Steinmann, E. E. Grafton-Cardwell and J. A. Rosenheim, An ecoinformatics approach to field scale evaluation of pesticide efficacy and hazards in California citrus, in prep.
[18] J. A. Rosenheim and C. Gratton, Ecoinformatics (big data) for agricultural entomology: Pitfalls, progress, and promise, *Annual Review of Entomology*, **62** (2017), 399–417.
[19] J. A. Rosenheim, S. Parsa, A. A. Forbes, W. A. Krimmel, Y. H. Law, M. Segoli, M. Segoli, F. S. Sivakoff, T. Zaviezo and K. Gross, Ecoinformatics for integrated pest management: Expanding the applied insect ecologist's tool-kit, *Journal of Economic Entomology*, **104** (2011), 331–342.

Received July 31, 2017; Accepted November 30, 2017.

E-mail address: htbanks@ncsu.edu
E-mail address: rarodge2@ncsu.edu
E-mail address: nmurad@ncsu.edu
E-mail address: rdwhite@ncsu.edu
E-mail address: jebanks@csumb.edu
E-mail address: bncass@ucdavis.edu
E-mail address: jarosenheim@ucdavis.edu