Comparison of Dynamic Models for Plant–Insect Herbivore–Pesticide Interactions

H.T. Banks, J.E. Banks, Jared Catenacci, Michele L. Joyner, and J.D. Stark

Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695
and
Undergraduate Research Opportunities Center (UROC)
California State University, Monterey Bay
Seaside, CA 93955
and
The Johns Hopkins University Applied Physics Laboratory
Laurel, MD 20723
and
Dept of Mathematics and Statistics
East Tennessee State University
Johnson City, TN 37614
and
Department of Entomology
Washington State University
Puyallup, WA 98371

July 30, 2019

Abstract

We consider a population dynamics model in investigating data from controlled experiments with aphids in broccoli patches surrounded by different margin types (bare or weedy ground) and three levels of insecticide spray (no, light, or heavy spray). The experimental data is clearly aggregate in nature. We compare two approaches, one of which ignores this aggregate nature and one which embraces this aspect of the experiments, to carry out parameter estimation computations along with statistical analysis to compare the two approaches using several model dynamics.

Keywords: Plant–insect interactions, inverse problems, hypothesis testing and standard errors in dynamical models, aggregate data, Prohorov metric

1 Introduction

In the summer of 1999, Banks and Stark [25] conducted a full-factorial design field experiment to explore the potentially combined effects of vegetation diversity and selective pesticide disturbance on aphid populations in a broccoli agroecosystem. They applied three concentration levels of the biorational pesticide imidacloprid to broccoli patches surrounded by either bare ground or weedy vegetation. These efforts by necessity resulted in what is now referred to as *aggregate data* [3, Chapter 14],[18, Chapter 5] wherein possibly (almost certainly) different subpopulations are counted in a longitudinal manner over the life of the experiment.

In this paper we describe our efforts fitting a population dynamics model to data from the Banks-Stark field study. In an earlier effort [2] we first used ordinary least squares techniques to fit several autonomous ordinary differential equation (ODE) models to each of the six data sets (two margin types each with three pesticide spray levels). Motivated by both varying environmental factors and changes in dynamics due to pesticide application, we also investigated the same set of models with piecewise constant and piecewise linear time-varying coefficients in the corresponding non-autonomous ODEs. Parameterizing these time-varying coefficients introduces additional degrees of freedom, but in general the non-autonomous systems provided substantially better fits to the data. Specifically, chi-squared tests revealed that increasing degrees of freedom to move from constant to time-varying coefficients yields statistically significant improvement in model fit.

We examined two of our models that characterize the population dynamics and compare model parameters under various margin types and spray levels. Our ultimate goal was to understand the influence of natural enemies or other margin-based factors on pest suppression separately from and in interaction with that of the insecticide. In these earlier efforts we treated the data as simple longitudinal data, ignoring the aggregate nature of the data sets. In the second part of our presentation here, we acknowledge the aggregate nature and apply the more recently developed methodology for treating such data sets in the context of the Prohorov Metric Framework (PMF) [3, 19, 18, 9, 20, 4, 8, 15, 21].

An obvious goal of that earlier paper and our current effort here is to contribute to the literature on plant–insect herbivore–pesticide interactions, thus raising questions for future investigations supported by experiments. In addition, we present a combined mathematical/statistical modelling methodology that has broad applicability to other problems in biology as well as engineering and general sciences.

2 Field study methods and data description

Banks and Stark conducted their full-factorial design field experiment [25] during the summer of 1999 at a Washington State University experimental farm in Puyallup, Washington. They established 2.5 meter square plots each containing 16 broccoli plants and surrounded by 1 meter wide margins of either naturally-occuring weedy vegetation or bare ground. At three points in the growing season, broccoli in each type of margin plot was treated with no pesticide spray, low concentration (15 g ai/ha; active ingredient per hectare) imidacloprid spray, or high concentration (30 g ai/ha) imidacloprid spray. Two replicates of each of the six treatment/margin combinations were placed in each of three fields (blocks) for a total of 36 experimental plots. Thus we have data from six plots for each of the following condition pairs:

1. Bare margin, no spray;
2. Bare margin, low spray (15 g ai/ha);
3. Bare margin, high spray (30 g ai/ha);
4. Weedy margin, no spray;
5. Weedy margin, low spray (15 g ai/ha); and
6. Weedy margin, high spray (30 g ai/ha).

Unwanted weedy vegetation within broccoli areas and between plots was regularly removed by a combination of tractor and hand cultivation. Plots were watered regularly throughout the growing season and dead or missing plants were replaced by similar sized plants as needed. The study commenced with transplantation in late June; pesticide spraying began in late July; and the study concluded in September. Imidacloprid spray was applied on July 23, August 13, and August 27, denoted by days 0, 21, and 35, respectively, in this paper.

At 4, 7, and 10 days after each pesticide spray (for a total of 9 times), Banks and Stark randomly selected a subset of 8 plants in each plot and visually censused the aphids. They counted all aphids on both sides of broccoli leaves and all other surfaces. Average cylindrical plant volume in each plot was obtained mid-season by measuring broccoli plant dimensions. Herbivore response to treatment manipulations was then calculated by dividing the number of aphids on a plant by the mean plant volume for that plot. Thus the census and volumetric measurements combined to yield a measure of aphid density (aphids per cubic meter) for each plot. In both Banks and Stark [25] and our present effort, inter-block variability was reduced by averaging the data across the six plots of each type (i.e., across three blocks, each with two replicates) to obtain a mean measure of aphid density over time. We fit our ODE models to this mean density data. The data sets can be viewed along with model fitting results in Section 4.2.

3 Mathematical models

For our efforts here, we only consider one type of ordinary differential equation model of population dynamics. While the authors of [2] considered 7 different models, we consider only one of these models to compare results when treating the data as truly aggregate data versus ignoring the aggregate nature of the data. This model was called Model 3 in the earlier reference [2]. The model involves a single state variable $N(t)$, which denotes the aphid population density: (mean aphids)/m^3. The model has the general form

$$\frac{dN(t)}{dt} = B(N(t),t)\,N(t) - D(N(t),t)\,N(t), \qquad \text{(MM)}$$

where $\frac{dN}{dt}$ (equivalently $\dot{N}$) denotes the time derivative of the state variable $N(t)$; B, a (potentially time- and/or state-dependent and/or random variable) birth rate; and D, a (potentially time- and/or state-dependent and/or random variable) death rate.

The model used in our current study (omitting the implicit time dependence of N and $\dot{N}$) assumes mortality due to pesticides and/or predation is the driving force behind aphid population dynamics and is the standard exponential model for population dynamics, allowing for simultaneous exponential birth and death. Here $B > 0$ and $D > 0$ in (MM) above for the deterministic time dependent coefficient models. In the case of random differential equation models we take $B - D = A$ for a random variable and we have an RDE model

$$\dot{N} = BN - DN = AN. \qquad (1)$$

Note that we let $A = B - D$ denote a combined birth/death rate in this model. Because aphids reproduce by parthenogenesis (i.e., asexually) in the field, exponential growth is a reasonable assumption for their population dynamics, especially when densities are below the carrying capacity. Densities during the field study considered here were consistently below carrying capacity densities observed in previous years.

For more information on models of the form (MM), see Boyce and DiPrima [27]. Note that in each case, a solution to the ordinary differential equation is uniquely determined by imposing a single initial condition, denoted $N_0 = N(t_0)$, where $t_0 = 0$ is the initial time considered in our studies.

4 Treatment as Simple Longitudinal Data as in [2]

Ignoring the aggregate nature of the data, we first fitted the models with a constant coefficient, a, to the data, and then repeated, allowing a time-varying coefficient, $a(t)$. In the time-varying case, we first employed piecewise constant and then piecewise linear coefficients. Figure 1 depicts a sample of the time varying coefficients considered.

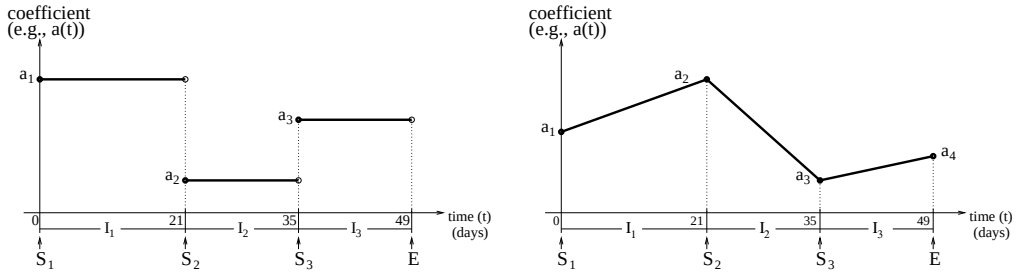

Figure 1: Example of time-varying coefficients. Left plot: piecewise constant, values a_i are on intervals; Right plot: piecewise linear, values a_i are at nodes. **S** denotes spray application and **E** denotes end of study.

Thus in [2] we used either $a(t) = \sum_{i=1}^{3} a_i \chi_{I_i}(t)$ or $a(t) = \sum_{i=1}^{4} a_i \phi_i(t)$, where $\chi_{I_i}(t)$ is the characteristic function which has value 1 on interval I_i and 0 elsewhere and the $\phi_i(t)$ are piecewise linear splines.

We expected better model fits (smaller residuals and better visual fits) using the non-autonomous models, due to the larger number of degrees of freedom. Certainly the residuals will be at least as small with time-varying coefficients as with constant coefficients. We use statistical analysis (described in detail in [2] and [18]) to determine whether any improvement in fit is strictly due to the increased degrees of freedom, or is statistically significant given the increase in degrees of freedom.

4.1 Least squares problem formulation and solution

We fit the model described in Section 3 to each of the six data sets, using an ordinary least squares cost functional to measure the model fit to the data. Details of the ordinary least squares inverse problems and the algorithms are given in [2].

The increased number of parameters in the models with a time-varying coefficient make them more difficult to fit. However, the optimal coefficients from the constant coefficient case and the results of the sampling algorithm prove good initial iterates for the least squares optimization to find piecewise defined coefficients in nearly all cases. We note that the optimal parameters found are not unique, but in many cases at least provide good fit to data and reasonable residuals.

The authors of [2] tested the significance of adding additional degrees of freedom to the inverse problem, i.e., using piecewise constant ($s = 2$ additional degrees of freedom) or piecewise linear ($s = 3$ additional degrees of freedom) coefficients, by comparing the cost function values at optimal parameters for constant coefficients (J_{cons}) and those using piecewise coefficients (J_{pw}). For each model, the authors of [2] compared the improvement in using the piecewise constant (pwc) or piecewise linear (pwl) coefficient (non-autonomous) model over the corresponding autonomous model. Using the theory developed by Banks and Fitzpatrick [12] and in particular, the test statistic

$$U = \frac{n\left[J_{cons}(q^*_{cons}) - J_{pw}(q^*_{pw})\right]}{J_{pw}(q^*_{pw})}, \tag{2}$$

where $n = 9$ data points, and q^*_{cons} and q^*_{pw} denotes either q^*_{pwc} or q^*_{pwl} optimal parameters for the two cases, in a χ^2 test with the null hypothesis that constant coefficients are sufficient to fit the data. By computing

the tail probability α beyond $U = U_9$ in a $\chi^2(s)$ distribution (s denotes number of additional degrees of freedom), one can determine the maximum level of confidence $P = (1 - \alpha)$ at which one can reject the null hypothesis. This allows the authors in many cases to suggest with confidence that the improvement in model fit achieved with time-varying coefficients was significant. Banks and Fitzpatrick [12] developed theoretical foundations and applied this method to similar problems in 1990 and additional examples can be found in [17]. We note that while the methodology developed in [12] uses an ANOVA-like test statistic (2), it is a more sophisticated model comparison technique designed to permit comparison of time-varying nonlinear model dynamics with a variety of mechanistic assumptions (e.g., pesticide / population / margin interaction mechanisms in the biological models treated in [2]) as illustrated in our model comparisons below. The test statistic depends explicitly on the number of observations n and implicitly on the difference s in the number of parameters being estimated through the degrees of freedom of the asymptotic limit χ^2.

Of course any estimates of model parameters from data should be accompanied by an estimate of the associated uncertainty. We use here (as did the authors in [2]) the standard methods based on sensitivity equations explained in detail in [18, Chapter 3] .

4.2 Model fitting results

We briefly report here on some of the estimated parameters and fits to data for Model 3 from [2] (Equation (MM) above). However, we restrict the results to only the constant and piecewise constant cases as these are the two cases we use for comparison when considering the aggregate nature of the data. We will refer the reader to [2] for results of the other models and the piecewise linear case.

In Figures 2 through 7 and Tables 1 through 6, we present a visual fit to data and estimated parameters with constant (a) and piecewise constant ($a(t)$) coefficients, on each of the six data sets. We include the standard errors σ for the constant case and the values for the cost functions and test statistic U in each case. We note that in all cases except for the weedy margin, low spray dataset (Figure 6 and Table 5), the piecewise constant provided a statistically better fit with a confidence level over .77 where the confidence is given by $P = 1 - \alpha$.

We found that our results when ignoring the aggregate nature of the data statistically justify the incorporation of time-dependent parameters in the models. Moreover, as shown in [2], we found little perceptible difference in statistical significance between adding piecewise constant versus piecewise linear coefficients. The form of these coefficients could of course be further refined with more information on the biological processes and environmental influences.

When using time-varying rates, our estimations for birth and death parameter values are within an order of magnitude of the values derived in an (*in vitro*) laboratory study (unpublished) by Stark. This represents a **good fit** in general between models and experimental data, especially since it incorporates observation and process error from field data as well as the difference between the open system in the field (*in vivo*) and the closed laboratory setting (*in vitro*) in which the Stark values were derived.

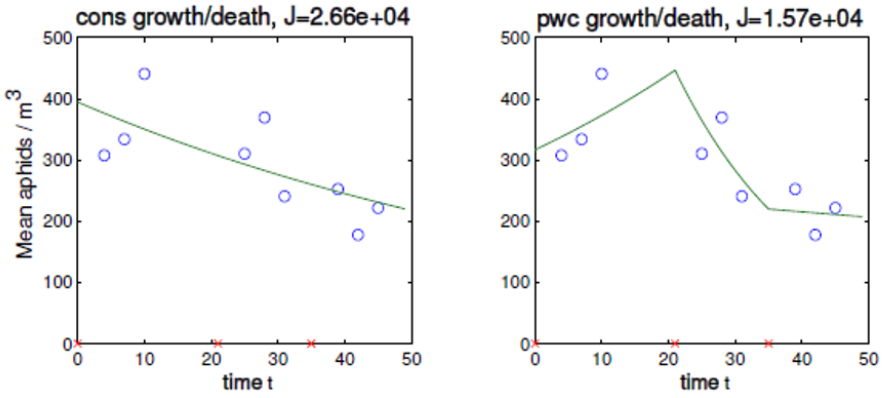

Figure 2: Fit of model with constant a and piecewise constant $a(t)$ to data from the **bare ground**, **no spray** situation.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$) var	μ	σ
cons	a	-1.1919e-02	4.7231e-03
	N_0	395	47
	J	2.6554e+04	
pwc $a(t)$	a_1	1.6415e-02	
	a_2	-5.0672e-02	
	a_3	-4.1650e-03	
	N_0	316	
	J	1.5651e+04	
	U	6.27	
	$1-\alpha$	0.956	

Table 1: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **bare ground** with **no spray**.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$) var	μ	σ
cons	a	-3.0353e-03	9.3249e-03
	N_0	181	48
	J	3.3400e+04	
pwc $a(t)$	a_1	1.3464e-02	
	a_2	5.7674e-03	
	a_3	-6.9067e-02	
	N_0	149	
	J	2.5052e+04	
	U	3.00	
	$1-\alpha$	0.777	

Table 2: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **bare ground** with **low spray**.

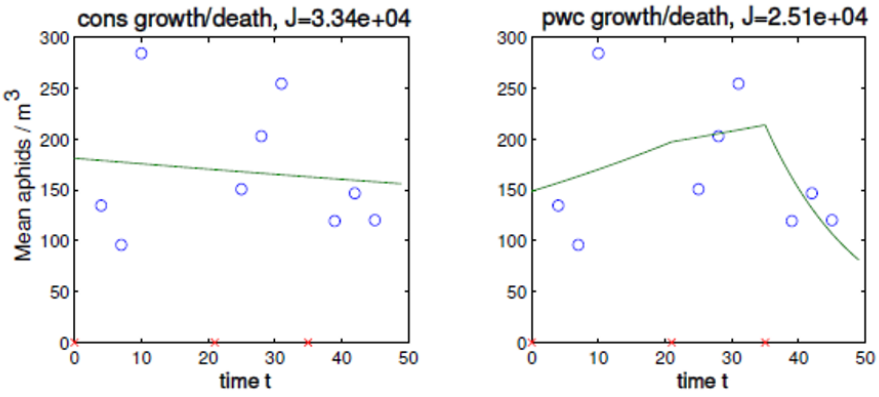

Figure 3: Fit of model with constant a and piecewise constant $a(t)$ to data from the **bare ground**, **low spray** situation.

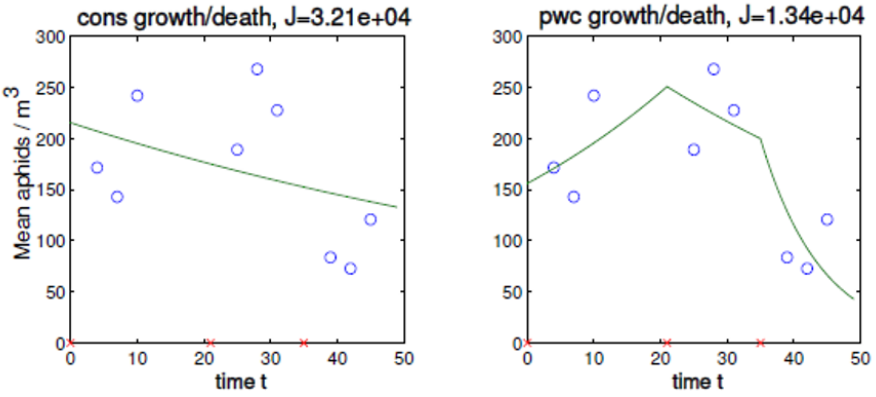

Figure 4: Fit of model with constant a and piecewise constant $a(t)$ to data from the **bare ground**, **high spray** situation.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$)		
	var	μ	σ
cons	a	-9.8693e-03	9.0441e-03
	N_0	216	51
	J	3.2146e+04	
pwc $a(t)$	a_1	2.2632e-02	
	a_2	-1.6214e-02	
	a_3	-1.1007e-01	
	N_0	156	
	J	1.3397e+04	
	U	12.59	
	$1-\alpha$	0.998	

Table 3: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **bare ground** with **high spray**.

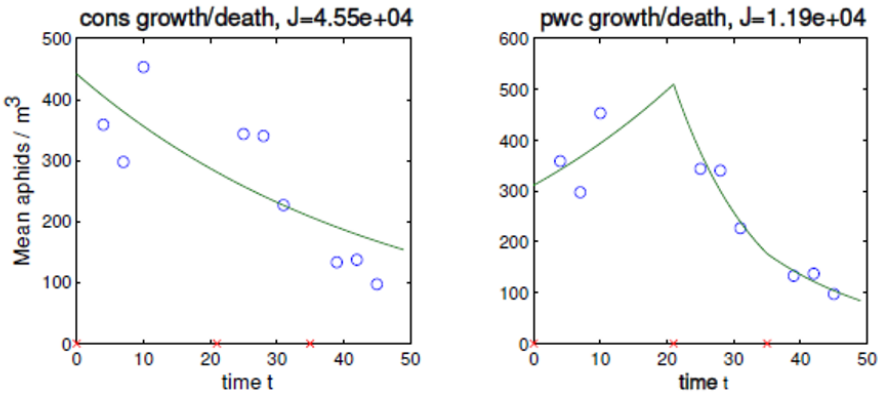

Figure 5: Fit of model with constant a and piecewise constant $a(t)$ to data from the **weedy ground**, **no spray** situation.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$) var	μ	σ
cons	a	-2.1539e-02	7.0853e-03
	N_0	442	67
	J	4.5489e+04	
pwc $a(t)$	a_1	2.3544e-02	
	a_2	-7.5684e-02	
	a_3	-5.2870e-02	
	N_0	311	
	J	1.1893e+04	
	U	25.42	
	$1-\alpha$	1.000	

Table 4: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **weedy ground** with **no spray**.

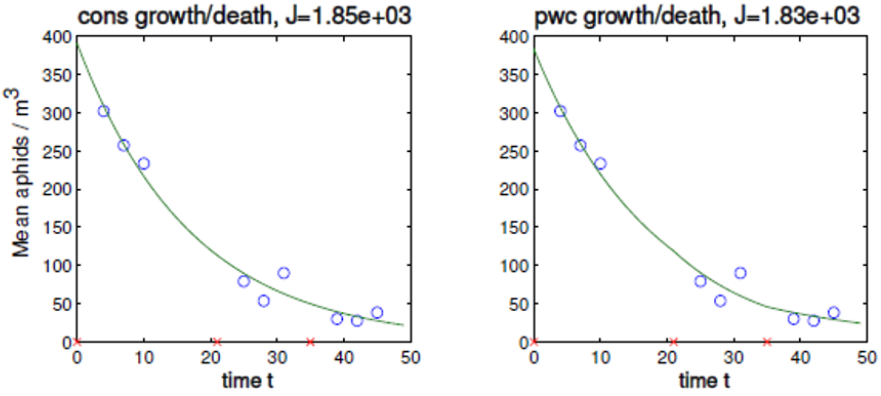

Figure 6: Fit of model with constant a and piecewise constant $a(t)$ to data from the **weedy ground**, **low spray** situation.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$) var	μ	σ
cons	a	-5.8838e-02	4.6498e-03
	N_0	391	20
	J	1.8532e+03	
pwc $a(t)$	a_1	-5.5741e-02	
	a_2	-6.8120e-02	
	a_3	-4.4933e-02	
	N_0	384	
	J	1.8312e+03	
	U	0.11	
	$1-\alpha$	0.053	

Table 5: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **weedy ground** with **low spray**.

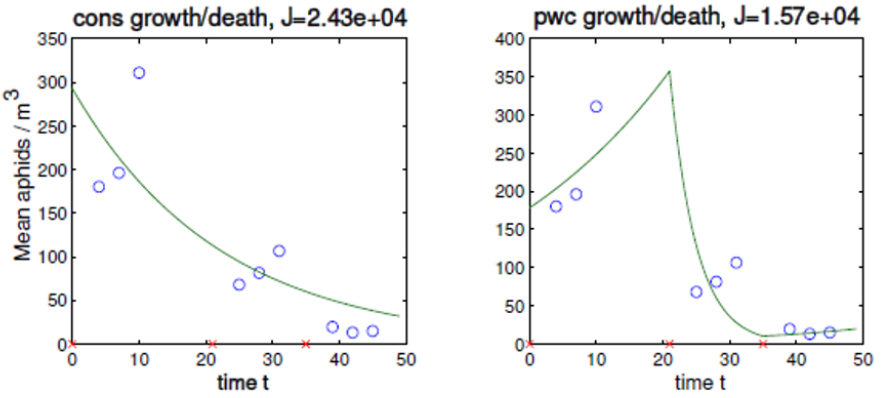

Figure 7: Fit of model with constant a and piecewise constant $a(t)$ to data from the **weedy ground**, **high spray** situation.

coeff type	Model ($\dot{N} = aN\ (a \in \mathbb{R})$) var	μ	σ
cons	a	-4.5308e-02	1.5274e-02
	N_0	293	63
	J	2.4298e+04	
pwc $a(t)$	a_1	3.3172e-02	
	a_2	-2.5439e-01	
	a_3	-4.7344e-02	
	N_0	178	
	J	1.5697e+04	
	U	4.93	
	$1-\alpha$	0.915	

Table 6: Estimated parameters μ and standard error σ, cost function J, test statistic U and confidence $P = 1 - \alpha$ for the model *ignoring the aggregate nature of the data* for data from **weedy ground** with **high spray**.

The dynamic models, parameter estimation techniques, and statistical methods discussed in [2, 18, 22] are tools mathematicians and field biologists might use to analyze time series population data. These tools are commonplace in the applied mathematical literature, but are not typically applied to data sets such as those considered in this study. We demonstrated the application of these mathematical and statistical methods to some example data sets and we propose their application to richer data sets, thus possibly yielding greater accuracy and applicability of the asymptotic statistical methods. The methods described can help us better understand the dynamics of complex interactions between living species and external factors and are readily extensible to other scenarios in the biological sciences.

However, one aspect of modeling that we did not address in [2] concerns random fluctuations in the populations. Specifically, where the data warrants it, one can (and should) consider variation of parameters across the population, i.e., parameter fluctuations in the population. This can be done with some type of stochasticity in the modeling such as use of stochastic differential equations or treatment of the parameters themselves as random across the population. However, the most grievous of errors in the above summary and in [2] is the lack of recognition of the aggregate nature of the data. We turn to this in the sections below.

5 Treatment as *aggregate* longitudinal data in the PMF

The modeling problem considered in this paper is referred to as an *individual dynamics / aggregate data problem* or a *Type I* problem in [18] in which we have a mathematical model for individual dynamics but only have aggregate data, not individual data. In this case, we have data from randomly selected plants in multiple plots which were averaged to reduce inter-block variability (*aggregate data*). Therefore, we should *not* be interested in a single parameter value a for our model; instead, the underlying goal should be to determine a probability measure which describes the distribution of the parameter across all members of the population.

In Section 4.2, the authors of [2] estimated the parameter a or $a(t)$ in a deterministic, least squares setting. In other words, it was assumed the experimental data $\{\hat{z}_i\}$, $i = 1, ..., n$ corresponds to individual observations of the state variable, i.e., z_i, $i = 1, ..., n$ are realizations of a random variable Z_i satisfying the statistical model

$$Z_i = N(t_i; a_0) + \mathcal{E}_i, \quad i = 1, ..., n$$

(when assuming constant variance) [2, 18]. Here $N(t; a_0)$ is the solution to (1) with some "truth" parameter a_0 and error $\mathcal{E}_i$. However, in the collection of the data in this paper, the experimental observations were collected from multiple plants and *averaged* across different plots. Therefore, we do not have longitudinal measurements of an independent state variable but expected values across the population which we will refer to as $\{\hat{u}_i\}$, $i = 1, ..., n$. If the parameters do not vary to a large extent across the population, then these "mean" value approximations from Section 4.2 may suffice. However, in many cases, these "mean" value approximations, in which the aggregate nature of the data is ignored, *may lead to inaccurate parameter estimates and an inaccurate description of the dynamics within the population*. Therefore, in this paper, we want to adequately consider the aggregate nature of the data.

To incorporate the potential population-dependent variability in the model parameter, we now assume the parameter A in Equation (1) is a random variable with probability distribution P which belongs to a probability space $P(\mathcal{Q})$ that may be infinite dimensional. As in [6], we define the set $P(\mathcal{Q})$ of all probability distributions on the admissible parameter space $\mathcal{Q}$ and seek to minimize the objective function

$$\begin{aligned} J(P) &= \sum_{i=1}^{n} |u(t_i; P) - \hat{u}_i|^2 \\ &= \sum_{i=1}^{n} \left((u(t_i; P))^2 - 2u(t_i; P)\hat{u}_i + (\hat{u}_i)^2 \right) \end{aligned} \tag{3}$$

over $P \in P(\mathcal{Q})$ where the expected value is defined by

$$u(t;P) = \mathcal{E}[N(t)|P] = \int_Q N(t;a)dP(a). \tag{4}$$

If it is possible to predict the expected form of the probability distribution P *a priori*, then we could use the parametric form of the distribution in (4) and estimate the parameters describing the distribution by minimizing (3) using a least-squares methodology (see [28]). However, in general, it is typically not plausible to predict the expected form of the distribution a priori. In this case, $P(\mathcal{Q})$ may be chosen as the space of *all* probability distributions. Thus we need non-parametric approaches that do not require any assumptions with respect to the form of the probability distribution. In [10], two approaches were compared in which $P(\mathcal{Q})$ was approximated by finite dimensional sets $P^M(\mathcal{Q})$, yielding computationally feasible schemes. Using the Prohorov metric [7, 26] and the theoretical framework in [6], if $\mathcal{Q}$ is compact, then we can define a family of finite dimensional approximation problems in which we are guaranteed convergence (in the Prohorov metric) of a minimizer of

$$\begin{aligned} J(P^M) &= \sum_{i=1}^{n} \left| u(t_i;P^M) - \hat{u}_i \right|^2 \\ &= \sum_{i=1}^{n} \left((u(t_i;P^M))^2 - 2u(t_i;P^M)\hat{u}_i + (\hat{u}_i)^2 \right) \end{aligned} \tag{5}$$

over the set P^M to a minimizer for the original problem, i.e., the minimization of (3).

One approach is to assume a discrete collection of admissible parameter values $Q^M = \{q_k^M\}_{k=0}^M$ and approximate $P(\mathcal{Q})$ with

$$P^M(\mathcal{Q}) = \left\{ P \in P(\mathcal{Q}) | P' = \sum_k p_k^M \delta_{q_k^M}, \sum_k p_k^M = 1 \right\} \tag{6}$$

where $\delta_{q_k^M}$ is the delta function with an atom at q_k^M. In this case, the population density is approximated by

$$u(t;\{p_k^M\}) = \sum_{k=0}^{M} N(t;a_k^M)p_k^M$$

where $N(t;a_k^M)$ is the subpopulation density from (1) with birth/death rate a_k^M. This is the delta function approximation method, denoted DEL(M), where M is the number of nodes or elements used in this approximation.

An alternative approach, and the one used in this paper, is to assume P is a continuous probability distribution on the parameter and use piecewise linear splines to approximate the distribution. In this methodology, we still assume a discrete collection of admissible parameter values $Q^M = \{q_k^M\}_{k=0}^M$ but now approximate $P(\mathcal{Q})$ with

$$P^M(\mathcal{Q}) = \left\{ P \in P(\mathcal{Q}) | P' = \sum_k b_k^M l_k^M(q), \sum_k b_k^M \int_Q l_k^M(q)dq = 1 \right\} \tag{7}$$

where the piecewise linear splines are represented by l_k^M. Using this approximation method, the population density from (4) is approximated by

$$u(t;\{b_k^M\}) = \sum_{k=0}^{M} b_k^M \int_Q N(t;a)l_k^M(a)da$$

where $p_k^M(a) = b_k^M l_k^M(a)$ is the probability density for individuals in subpopulation k. This spline based method is denoted by SPL(M,N) where M is the number of basis elements used to approximate the birth/death rate probability distribution and N is the number of quadrature nodes used to approximate the integral.

In both approaches, the least squares problem in (5) reduces to a constrained quadratic programming problem [13, 14] in which we minimize

$$F(\mathbf{p}) = \mathbf{p}^T \mathbf{H} \mathbf{p} + 2\mathbf{p}^T \mathbf{f} + c \tag{8}$$

over $P^M(\mathcal{Q})$, where $\mathbf{p}$ is the vector containing p_k^M, $0 \leq k \leq M$ or b_k^M, $0 \leq k \leq M$ when using DEL(M) or SPL(M,N), respectively. We let $\mathbf{H}$ be the matrix with entries given by

$$H_{km} = \sum_i N(t_i; a_k^M) N(t_i; a_m^M);$$

$\mathbf{f}$ is the vector with entries

$$f_k = -\sum_i \hat{u}_i N(t_i; a_k^M)$$

and

$$c = \sum_i (\hat{u}_i)^2.$$

When using DEL(M), the constraint

$$\sum_{k=1}^{M} p_k^M = 1$$

is imposed; whereas when using SPL(M,N), we include the constraint

$$\sum_{k=1}^{M} b_k^M \int_Q l_k^M(q) dq = 1.$$

In [11], Banks and Davis provided a computational framework for the quantification of uncertainty associated with estimating probability distributions or densities for a Type 1, individual dynamics/aggregate data problem. In their trials, they computed confidence bands for both the cumulative distribution and probability density functions for the estimated probability distribution for a growth rate parameter in the Sinko-Streifer model. They were able to show that as the number of nodes, M, increases, the confidence bands appear to converge nicely until M becomes too large, i.e., the problem becomes over-parameterized. They showed that the condition number for the matrix $\mathbf{H}$ in the quadratic programming problem (8) increases as M increases and causes the inverse problem to become ill-posed.

Prior to estimating a probability distribution for the growth/death rate A in (1) using the experimental data, we performed trials using simulated data with a known probability distribution, assuming a small number of data points (as given in this study). Even for small values of M, the condition number was fairly large, providing poor fits to the true density. To improve the fit to the simulated data, we first incorporated regularization with a small regularization parameter and then determined that M=15 nodes provided a good fit to both the cumulative distribution and probability density functions for the known probability distribution as well as a good fit to the solution.

5.1 Model fitting results

In this section, we use the techniques outlined in Section 5 to estimate a probability distribution for random variable parameter A in Equation (1), accounting for the fact that we have aggregate data. We do this in two

ways; we first assumed that the distribution did not change with respect to time. This is comparable to the case in which a was assumed to be constant in Section 4.2; however, in this section we take into account the potential population-dependent variability in the model parameter since we have aggregate data. Therefore, in our initial efforts, we assume there is variability across the population, but there is no change in variability with respect to time. Hence, we estimate one probability distribution for A for the entire data set. The model fits to the solution for each of the six data sets are given in the left-hand plots in Figures 8, 11, 14, 17, 20, and 23 for bare ground - no spray, bare ground - low spray, bare ground - high spray, weedy ground - no spray, weedy ground - low spray and weedy ground - high spray data, respectively. This is comparable to the fits we saw in Section 4.2 in the left-hand plots of Figures 2- 7. The corresponding estimated probability density function and cumulative distribution function are given in Figures 9, 12, 15, 18, 21, and 24. We observe that the model fits to the data when considering the aggregate nature of the data and not assuming time dependence are nearly identical to the previous fits when the aggregate nature of the data was ignored. We note that the probability density function varies depending on the choice of $\mathcal{Q}^M$, i.e, it changes depending on the nodes which are chosen in the interval as well as the interval over which the nodes are chosen. This makes sense as one could miss crucial variability in the probability density function if nodes are not chosen for values in which there is a noticeable change. The choice of $\mathcal{Q}^M$ and the placement of the nodes is a future research question to be addressed. We note that in our initial trials, we chose $\mathcal{Q}^M$ as indicated by the standard error estimates for a given in Tables 1 through 6. However, using this large interval with only M=15 nodes, the estimates indicated point distributions at values close to the estimated constant values in Tables 1 through 6. This could explain the similarity in model fits and cost estimates between the two methodologies. Therefore, in this case, when ignoring time-dependence, the "mean" value approximation may suffice.

In our next efforts, we assumed time-dependence variability. In this situation, we split the time interval into three subintervals over which we estimated subinterval population distributions. In other words, we assumed on each subinterval, there was potentially a different probability distribution for A_i, $i = 1,\ 2,\ 3$. We considered three subintervals based on the previous efforts in [2] and due to the collection of the data between sprays. This scenario is comparable to the piecewise constant estimates discussed in Section 4.2. We assumed the solution was a *piecewise continuous function* in which the value of the solution at the end of one subinterval was the initial value of the solution at the start of the next subinterval. We want to emphasize the fact that these solutions are *not* unique. In future research efforts, one might like to investigate how one "best" chooses the interval endpoints as well as the choice of $\mathcal{Q}^M$ on each subinterval. In current efforts, the model fits can vary depending on these factors. The continuous piecewise model solution fits are given in the right plots in Figures 8, 11, 14, 17, 20, and 23 for bare ground - no spray, bare ground - low spray, bare ground - high spray, weedy ground - no spray, weedy ground - low spray and weedy ground - high spray data, respectively. These can be compared to the case when the aggregate nature of the data was ignored (see right-hand plots in Figures 2- 7). The estimated distributions on each subinterval are given in Figures 10, 13, 16, 19, 22, and 25. We note that when incorporating the time-dependence in the model, the model fits when incorporating the aggregate nature of the data resulted in lower cost functions than when the aggregate nature of the data was ignored in all cases except for the weedy ground, low spray data set. These are not nested models; therefore, we cannot statistically compare the model fits from Section 4.2 to the model fits in this section using the methodology outlined in Section 4.1. Nonetheless, the differences in model fits illustrates the potential errors which may occur when one naively treats aggregate longitudinal data as individual longitudinal data.

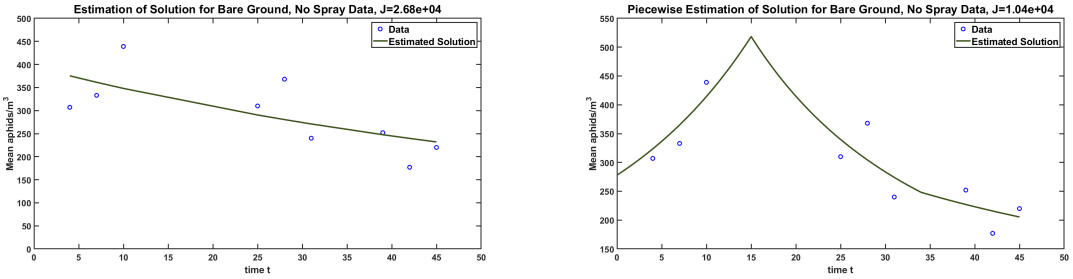

Figure 8: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **bare ground**, **no spray** situation.

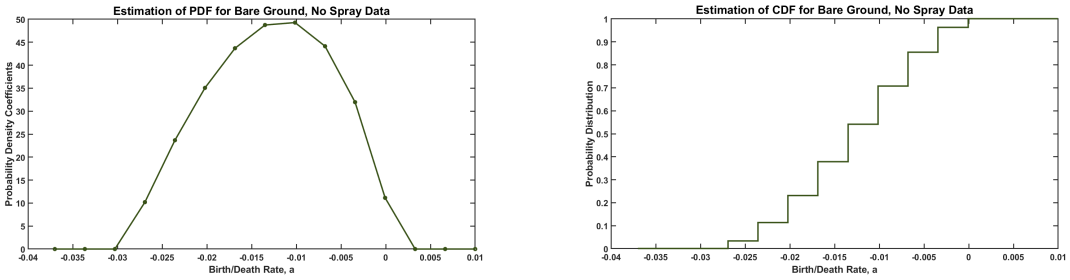

Figure 9: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **bare ground**, **no spray** situation when considering the *entire* data set.

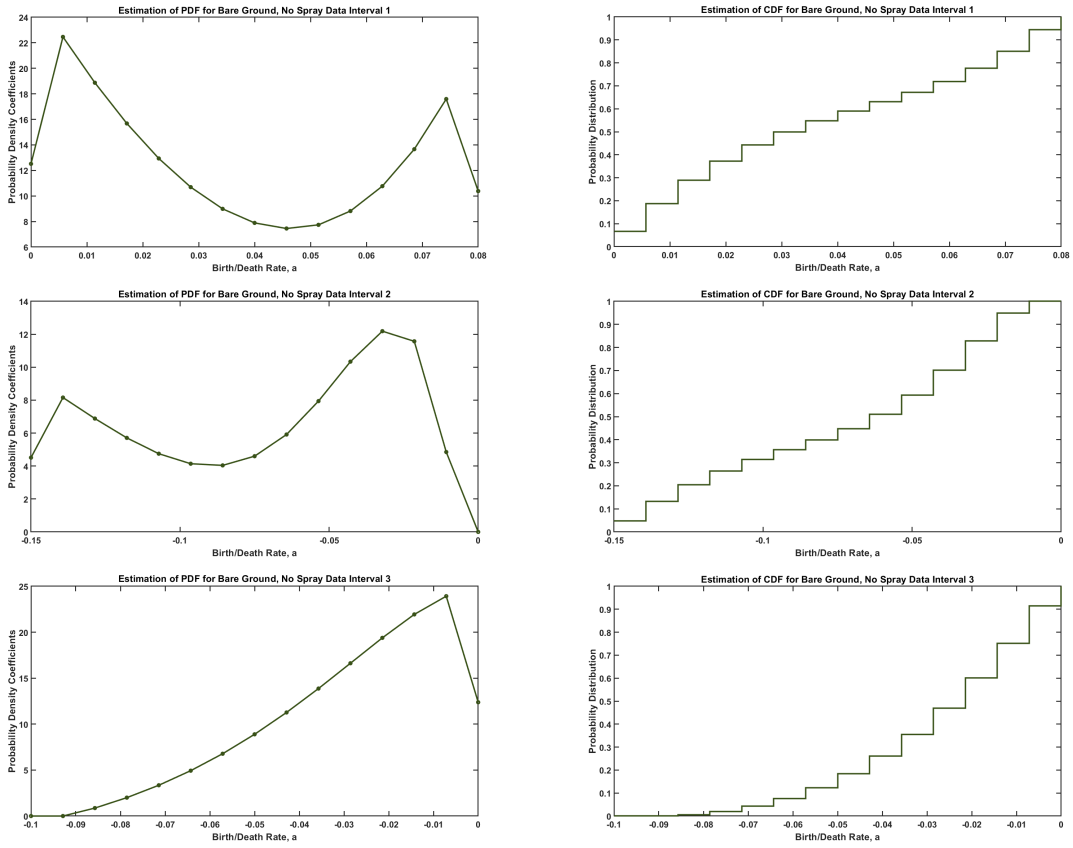

Figure 10: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **bare ground**, **no spray** situation.

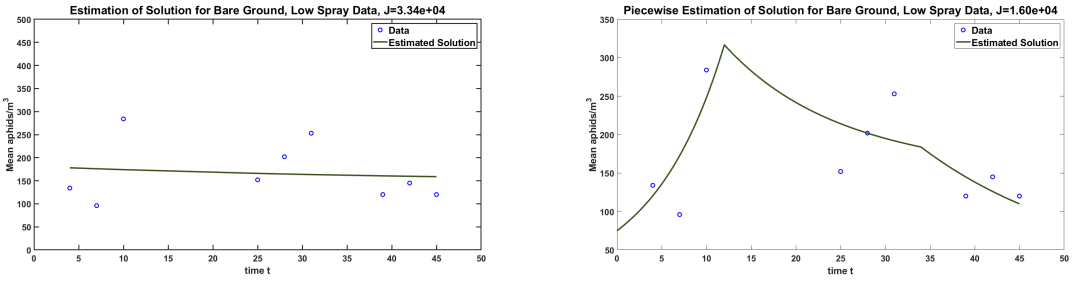

Figure 11: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **bare ground**, **low spray** situation.

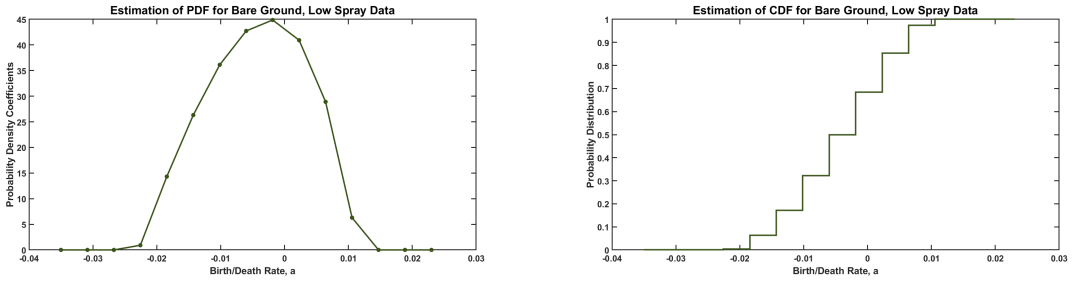

Figure 12: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **bare ground**, **low spray** situation when considering the *entire* data set.

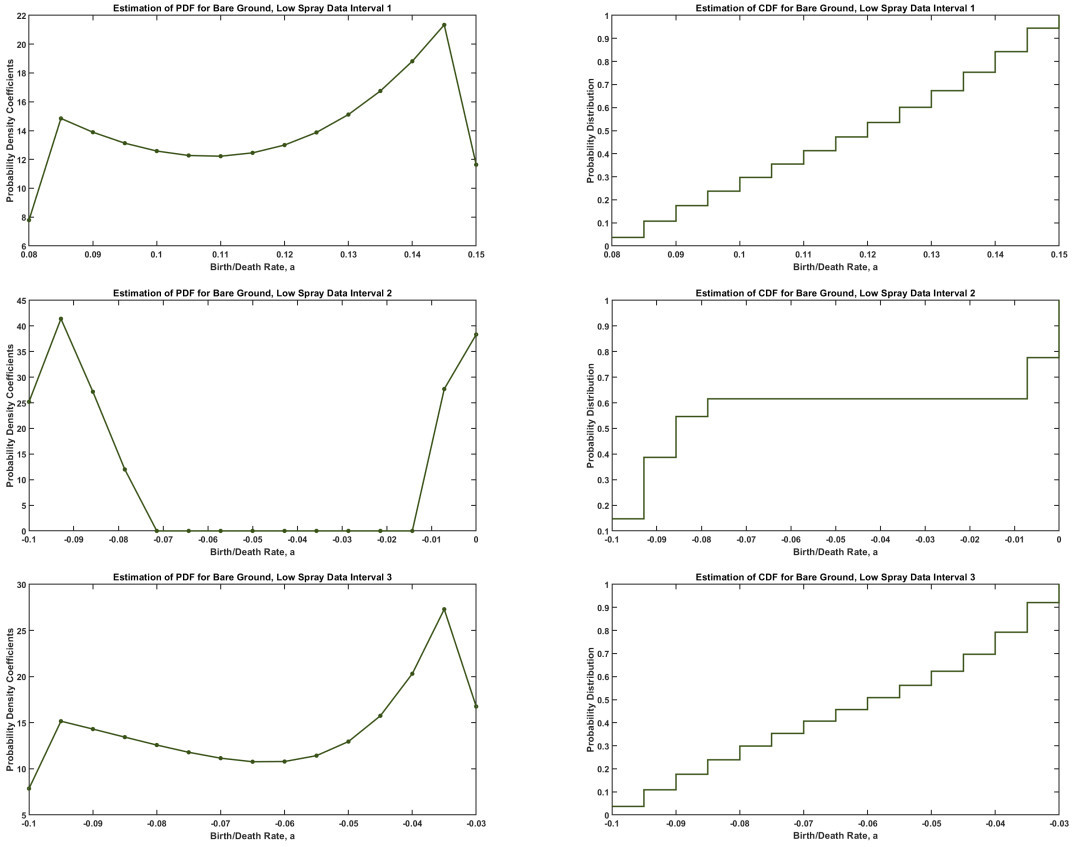

Figure 13: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **bare ground**, **low spray** situation.

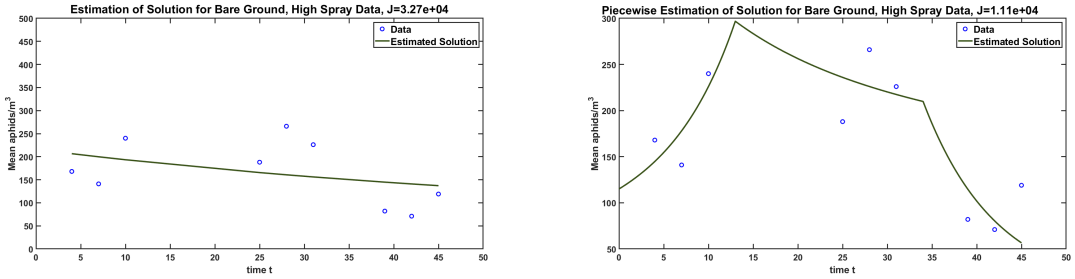

Figure 14: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **bare ground**, **high spray** situation.

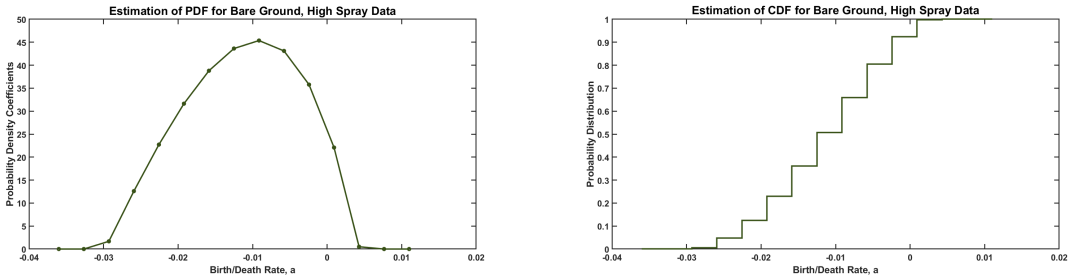

Figure 15: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **bare ground**, **high spray** situation when considering the *entire* data set.

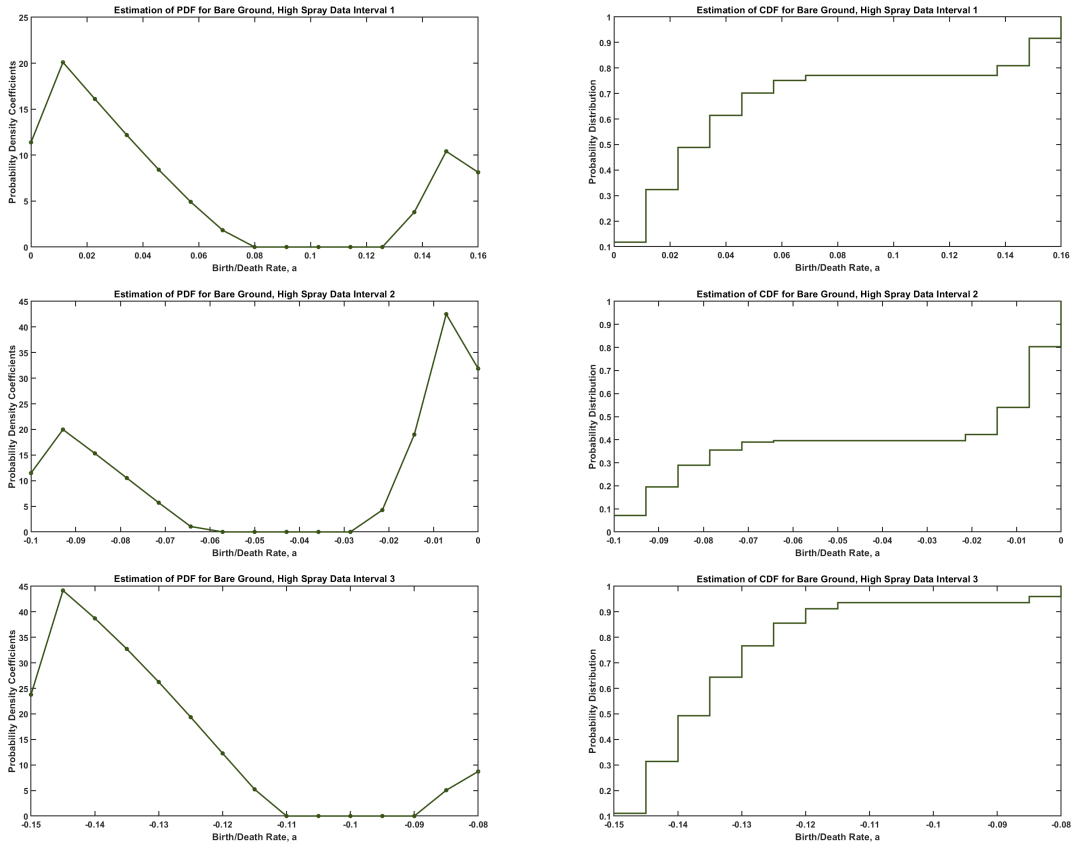

Figure 16: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **bare ground**, **high spray** situation.

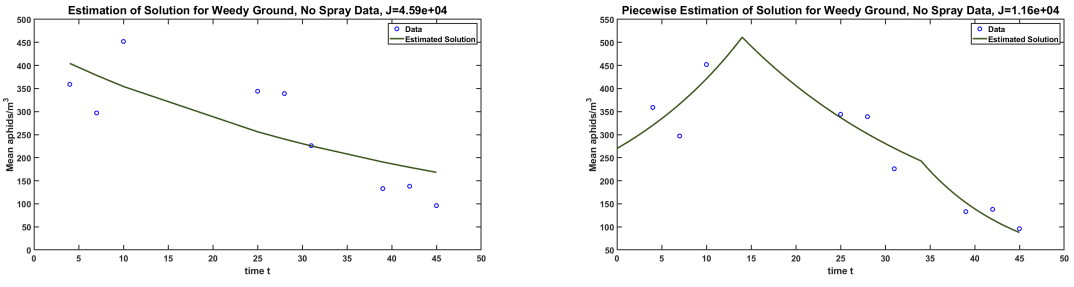

Figure 17: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **weedy ground**, **no spray** situation.

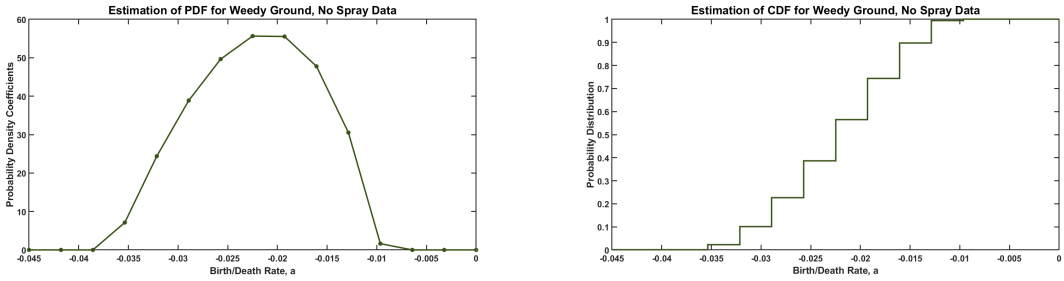

Figure 18: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **weedy ground**, **no spray** situation when considering the *entire* data set.

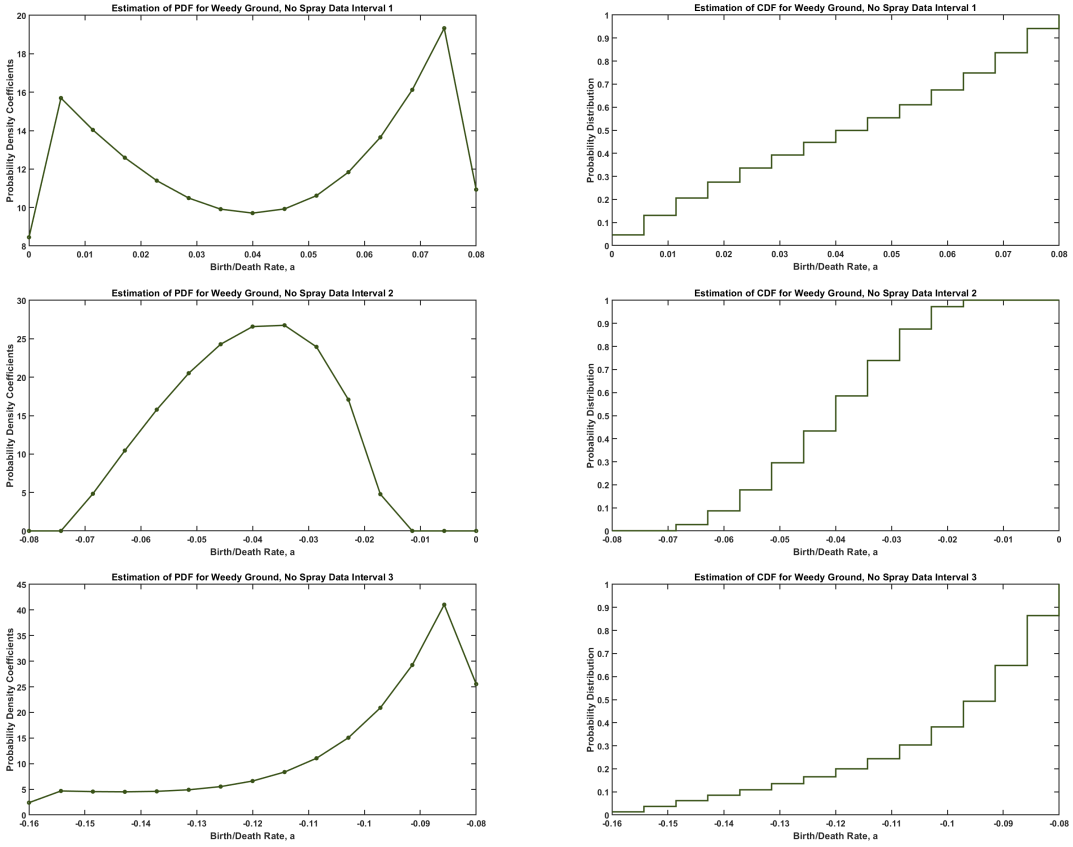

Figure 19: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **weedy ground**, **no spray** situation.

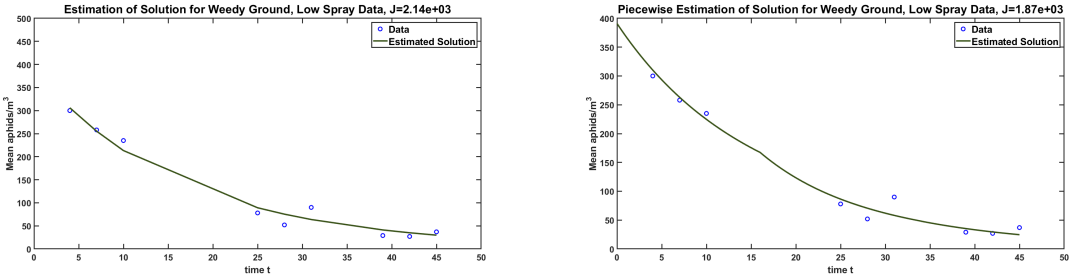

Figure 20: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **weedy ground**, **low spray** situation.

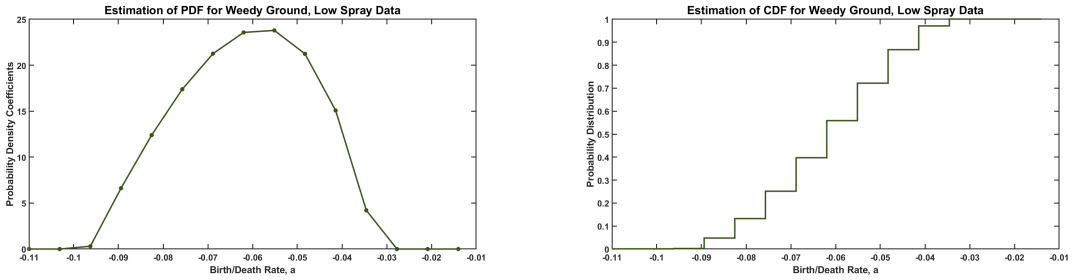

Figure 21: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **weedy ground**, **low spray** situation when considering the *entire* data set.

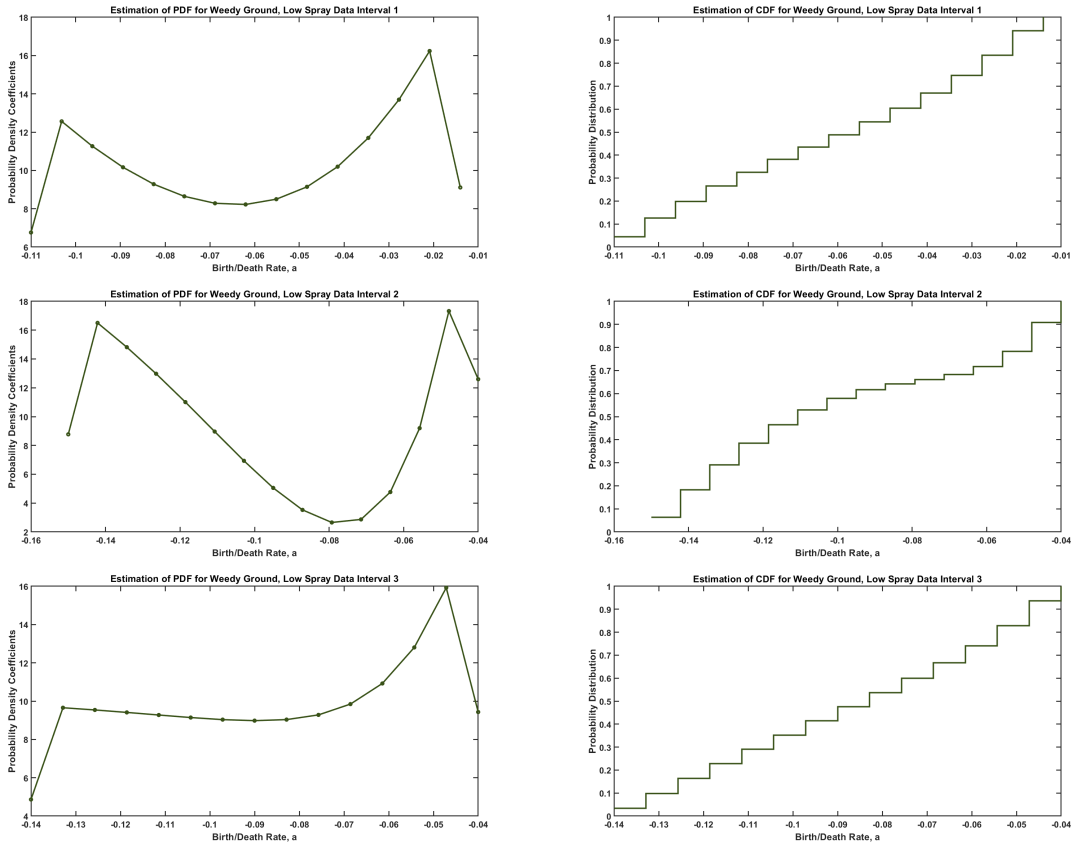

Figure 22: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **weedy ground**, **low spray** situation.

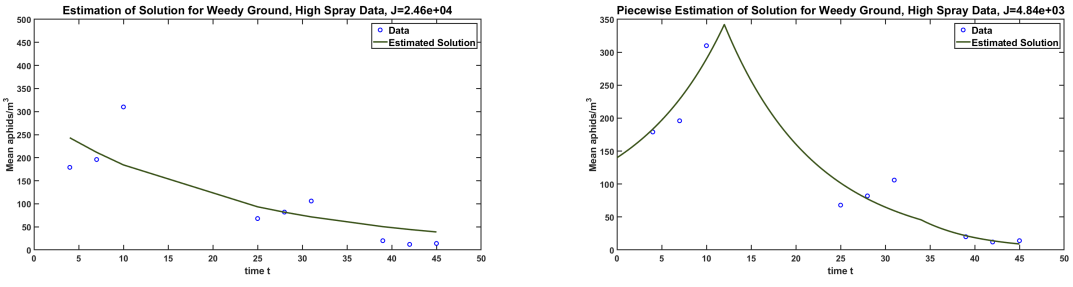

Figure 23: Fit of model with *distribution* and a *piecewise distribution* for the parameter A to data from the **weedy ground**, **high spray** situation.

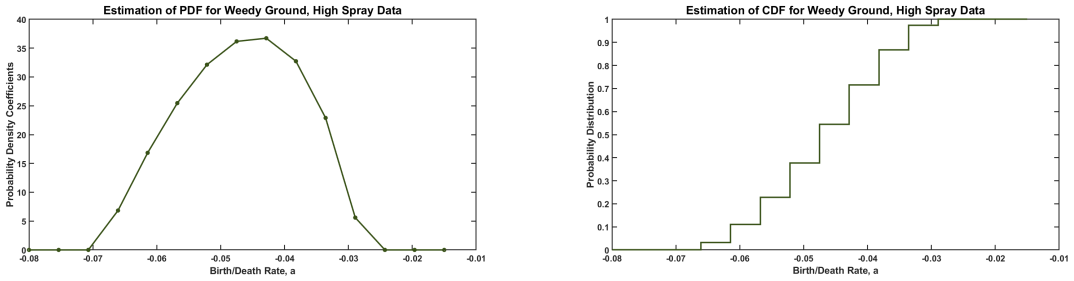

Figure 24: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter A to data from the **weedy ground**, **high spray** situation when considering the *entire* data set.

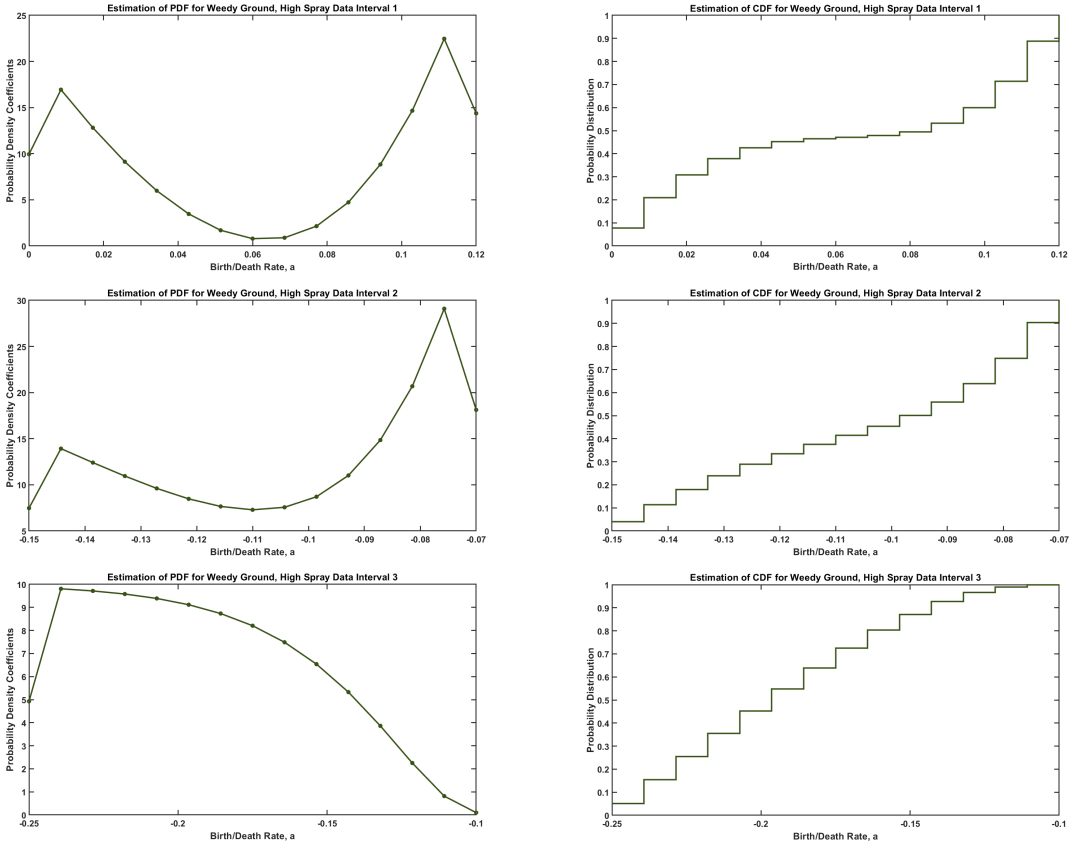

Figure 25: The estimated probability density function and cumulative distribution function for the probability *distribution* of the parameter $A(t)$ on each of the intervals when considering a *piecewise function* using 3 different intervals for data from the **weedy ground**, **high spray** situation.

6 Conclusions

One may on initial observation, perceive strong similarities in our findings above whether one treats the data as individual longitudinal data or as the aggregate data which it is. However, in the 3 subinterval approach shown above, the results were different with quite different cost functions. Moreover, in other examples not given in this paper, we found vastly different or non-converging estimates when treating the data *correctly* as aggregate data. Our results in this paper underscore the potential pitfalls inherent in using field census data in modeling population dynamics in applied ecology. Many sampling schemes in agroecosystems as well as natural systems rely on regular population counts. These data are often regarded as individual longitudinal data, which is reflected in assumptions underlying the statistical analyses. In many of these cases, however, because of reproduction, mortality, and redistribution of individuals, repeated population sampling results in counts of different individuals each time period. This necessitates treating the data as aggregate individual data. In the current illustration, the difference in data interpretation differs little between these two approaches; we offer here a strong cautionary tale nonetheless because in many cases the two approaches may lead to markedly different conclusions.

Acknowledgments

This research was supported in part by the U.S. Air Force Office of Scientific Research under grant AFOSR FA9550-18-1-0457.

References

[1] B.M. Adams and H.T. Banks and J.E. Banks and J.D. Stark, Population Dynamics Models in Plant–Insect Herbivore–Pesticide Interactions, Center for Research in Scientific Computation, North Carolina State University, CRSC-TR03-12, March, 2003; Revised August, 2003.

[2] B.M. Adams, H.T. Banks, J.E. Banks and J.D. Stark, Population dynamics models in plant-insect-herbivore-pesticide interactions, *Math. Biosci.*, **196** (2005), 39–64.

[3] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*, Taylor and Frances Publishing, CRC Press, Boca Raton, FL, 2012.

[4] H.T. Banks, J.E. Banks, Neha 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, p.1–12, 2017, Springer : DOI: 10.1007/978-3-319-55795-38.

[5] H.T. Banks, J.E. Banks, J. Rosenheim, and K. Tillman, Modeling populations of *Lygus Hesperus* on cotton fields in the San Joaquin Valley of California: The importance of statistical and mathematical model choice, CRSC-TR15-04, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, May, 2015; *J. Biological Dynamics*, **11** (2017), 25–39 DOI: 10.1080/17513758.2016.1143533

[6] H.T. Banks and K.L. Bihari, Modeling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111.

[7] H.T. Banks, D.M. Bortz, G.A. Pinter and L.K. Potter, Modeling and imaging techniques with potential for application in bioterrorism, Chapter 6 in *Bioterrorism: Mathematical Modeling Applications in Homeland Security*, (H.T. Banks and C. Castillo-Chavez, eds.), Frontiers in Applied Math, **FR28**, SIAM, 2003, Philadelphia, PA, 129—154.

[8] H.T. Banks and Jared Catenacci, Aggregate data and the Prohorov Metric Framework: Efficient gradient computation, CRSC-TR15-13, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, November, 2015; *Applied Mathematical Letters*, **56**, June 2016, 1—9.

[9] H.T. Banks, Jared Catenacci and Shuhua Hu, Asymptotic properties of probability measure estimators in a nonparametric model, CRSC TR14-05, N. C. State University, Raleigh, NC, May, 2014; *SIAM/ASA Journal on Uncertainty Quantification*, **3** (2015), 417–433.

[10] H.T. Banks and Jimena L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Applied Numerical Mathematics*, **57** (2007), 753–777.

[11] H.T. Banks and Jimena L. Davis, Quantifying uncertainty in the estimation of probability distributions, *Mathematical Biosciences & Engineering*, **5**, (2008), 647–667.

[12] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems,*J. Math. Biol.*, **28** (1990), 501–527.

[13] H.T. Banks and B.G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, *Quarterly of Applied Mathematics*, **49** (1991), 215–235.

[14] H.T. Banks, B.G. Fitzpatrick, L.K. Potter, and Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data, In *Stochastic Analysis, Control, Optimization and Applications*, (W. McEneaney, G. Yin and Q. Zhang, eds.), Birkhauser, 1989, Boston.

[15] H.T. Banks, K.B. Flores, I.G. Rosen, E.M. Rutter, Melike Sirlanci, and W. Clayton Thompson, The Prohorov Metric Framework and aggregate data inverse problems for random PDEs, CRSC-TR18-05, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, June, 2018; *Communications in Applied Analysis*, **22**, No. 3 (2018), 415–446.

[16] H.T. Banks and Michele L. Joyner, AIC under the framework of least squares estimation, CRSC-TR17-09, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, May, 2017; *Applied Math Letters,* **74**, December (2017), 33–45.

[17] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser,Boston, 1989.

[18] H.T. Banks, Shuhua Hu, and W. Clayton Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, Taylor and Frances Publishing, CRC Press, Boca Raton, FL, 2014.

[19] H.T. Banks and W. Clayton Thompson, Least squares estimation of probability measures in the Prohorov Metric Framework, CRSC-TR12-21, N. C. State University, Raleigh, NC, November, 2012.

[20] H.T. Banks and W. Clayton Thompson, Existence and consistency of a nonparametric estimator of probability measures in the Prohorov metric framework, *Intl. J. Pure and Applied Mathematics*, **103** (2015), 819–843; DOI:10.12732/ijpam.v103i4.15

[21] H.T. Banks and W. Clayton Thompson, Random delay differential equations and inverse problems for aggregate data problems, CRSC-TR18-07, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, July, 2018; *Eurasian J. Mathematical and Computer Applications*, **6**, No. 4 (2018), 4–16.

[22] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton London New York, 2009.

[23] J.E. Banks, The scale of landscape fragmentation affects herbivore response to vegetation heterogeneity, *Oecologia*, **117** (1998), 239–246.

[24] J.E. Banks, Effects of weedy field margins on *Myzus persicae* (Hemiptera: Aphididae) in a broccoli agroecosystem , *Pan-Pac. Entomol.*, **76** (2000), 95–101.

[25] J.E. Banks and J.D. Stark, Aphid response to vegetation diversity and insecticide disturbance, *Agric. Ecosyst. Environ.***103** (2004), 595–599.

[26] P. Billingsley, *Convergence of Probability Measures*, Wiley, 1968, New York.

[27] W.E. Boyce and R.C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, John Wiley and Sons,Inc., 1997, 6th edition, New York.

[28] L. K. Potter, Physiologically based pharmacokinetic models for the systemic transport of trichloroethylene, Ph. D. thesis, North Carolina State University, 2001, `www.lib.ncsu.edu`.

[29] J.D. Stark and J.E. Banks, Selective pesticides: are they less hazardous to the environment? *Bioscience*, **51** (2001), 980–982.