

# Experimental Design and Estimation of Growth Rate Distributions in Size-Structured Shrimp Populations

H.T. Banks, Jimena L. Davis, Stacey L. Ernstberger and Shuhua Hu  
Center for Research in Scientific Computation  
Center for Quantitative Sciences in Biomedicine  
North Carolina State University  
Raleigh, North Carolina 27695-8212

Elena Artimovich and Arun K. Dhar  
Advanced BioNutrition Corporation  
7155 Columbia Gateway Dr.  
Columbia, MD 21046

November 15, 2008

## Abstract

We discuss inverse problem results for problems involving the estimation of probability distributions using aggregate data for growth in populations. We begin with a mathematical model describing variability in the early growth process of size-structured shrimp populations and discuss a computational methodology for the design of experiments to validate the model and estimate growth rate distributions in shrimp populations. Parameter estimation findings using experimental data from experiments so designed for shrimp populations cultivated at Advanced BioNutrition Corporation are presented illustrating the usefulness of mathematical and statistical modeling in understanding the uncertainty in the growth dynamics of such populations.

**Key words:** Inverse problems, size-structured population models, growth rate distributions, mathematical and statistical modeling, uncertainty in growth dynamics

# 1 Introduction

In this paper we present results for an inverse problem for the estimation of growth rate distributions in size-structured shrimp populations. The estimation and inverse problem efforts discussed here are motivated by previous work on a hybrid model of the shrimp biomass/countermeasure production system developed and discussed in detail in [5]. One of the goals of this joint project with Advanced BioNutrition Corporation (ABN) was the development of a model for a system where one uses shrimp as a scaffold organism to produce large amounts of vaccine in response to a specific bio-toxic antigen challenge. Because the output of the biomass growth model will serve as input to a vaccine production model, the ability to accurately model the dynamics (including any uncertainty) of the size-structured shrimp population is critical.

We discuss inverse problem methodologies for growth rate distributions in population models involving uncertainty related to the data or observations employed. However, the problems are not only the usual statistical inverse problems arising wherein measurement errors provide the source of data uncertainty and hence parameter uncertainty. Rather, the *aggregate* nature of the data produces an additional need for uncertainty in the model itself. In these problems one has individual dynamics (growth rate differential equations) but only aggregate or population level longitudinal measurements without data on individuals. Thus the usual hierarchical/mixing distributions methods from statistics [20] are not applicable. The problems require inverse problem methodologies for estimation of probability distributions for growth at the population level along with uncertainty quantities (e.g., confidence bands or regions that play the role of the usual confidence intervals for Euclidean parameters) for the estimated parameters and their associated sampling distributions. The resulting inverse problems are thus ones involving estimation of probability distributions (or associated densities) for growth rates across a family of growth rates admissible in the population.

The outline of the paper is as follows. We begin (Section 2) with a review of a class of size-structured population models along with modifications to permit growth rate variability or uncertainty. These include growth rate distribution (GRD) models as well as those involving a Markov stochastic process for size. In Section 3 we outline an approach based on simulated inverse problems to aid in the design of experiments to produce data adequate for validation of the GRD models. We do this in the context of an assumed parametric family (truncated Gaussian) of distributions or densities and use quantities from asymptotic statistics to evaluate efficacy. Among questions investigated are sampling frequency and sampling size. In Section 4 we present the results for the parameter estimation problem for the growth rate distribution of shrimp populations using experimental data collected according to the designs resulting from the studies of Section 3. Concluding remarks are given in Section 5.

## 2 Size-Structured Population Models

The Sinko-Streifer (SS) model [32] and its variations have been widely used to describe various age and size-structured populations (see [2, 3, 7, 12, 13, 16, 25, 28] for only a sample). More recently, extensions of these models have been employed in cell population models where size is replaced by intensity of a label or marker [27].

One version of the SS model is given by

$$\begin{aligned} v_t(x, t; g) + (g(x, t)v(x, t; g))_x &= -m(x, t)v(x, t; g) \\ g(\underline{x})v(\underline{x}, t; g) &= \int_{\underline{x}}^{\bar{x}} \beta(x, t)v(x, t; g)dx \\ v(x, 0; g) &= v_0(x; g). \end{aligned} \tag{2.1}$$

We note that the population density of individuals with growth rate  $g$  at time  $t$  with size  $x$  is denoted by  $v(x, t; g)$ . The mortality rate of individuals in the population is represented by  $m(x, t)$ , while  $\beta(x, t)$  represents the fecundity or reproduction rate. The parameters  $\underline{x}$  and  $\bar{x}$  represent the minimum and maximum sizes of the individuals, respectively. The initial size density of the population at  $t = 0$  is given by  $v_0(x; g)$ . At any time  $t$ , the number of individuals in the population with sizes between  $\underline{x}$  and  $\bar{x}$  is given by

$$N_{SS}(t) = \int_{\underline{x}}^{\bar{x}} v(x, t; g)dx.$$

In the SS model the growth rate of all individuals in the population is given by the following deterministic growth model

$$\frac{dx}{dt} = g(x, t). \tag{2.2}$$

Therefore, all individuals of the same size at the same time are assumed to grow at the same rate  $g(x, t)$ . However, this assumption leads to a model that is incapable of predicting the dispersion and bifurcation evident in mosquitofish population data collected in rice fields [7, 12, 13]. Dispersion in size has also been observed in experimental data shown in Figures 1 and 2 for the early growth of shrimp collected from two different raceways at the Shrimp Mariculture Research Facility, Texas Agricultural Experiment Station in Corpus Christi, TX. Although the initial sizes of the shrimp were very similar, a great deal of variability was observed in the aggregate type longitudinal data as time progressed. Therefore, a reasonable model for this population should account for the variability in the size distribution data which is perhaps a result of variability in the individual growth rates across the population [19].

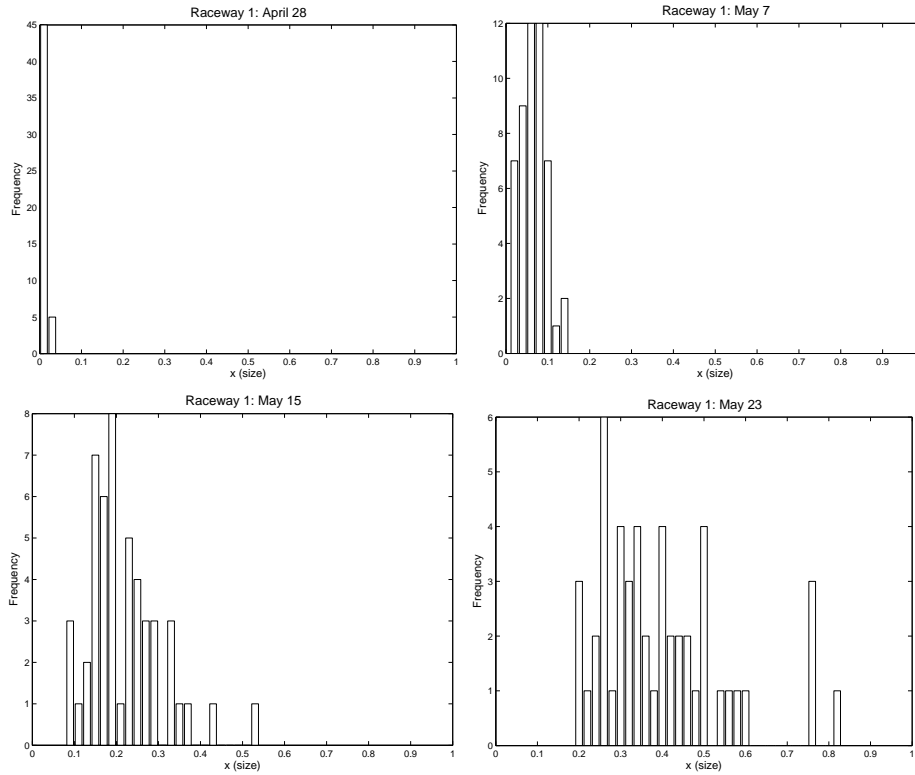


Figure 1: Histograms for longitudinal data for size (in grams) for Raceway 1.

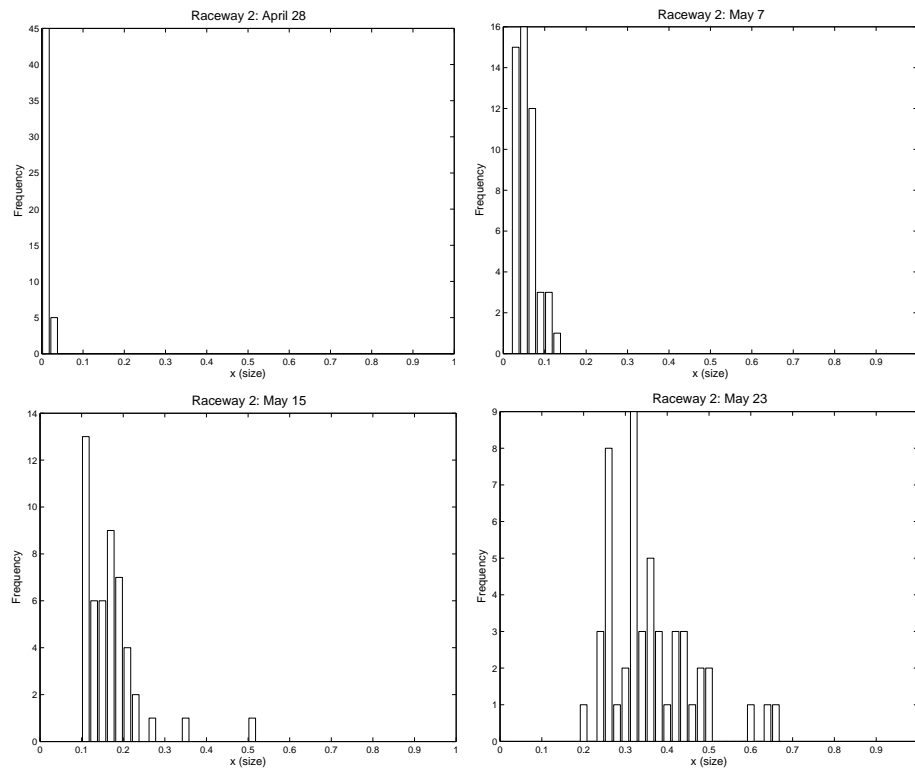


Figure 2: Histograms for longitudinal data for size (in grams) for Raceway 2.

In a modification of the SS model, the Growth Rate Distribution (GRD) model developed and investigated in [7, 12, 13], the growth rates are allowed to vary across the population, and individuals with the same growth rate are assumed to be in the same subpopulation. The population density in the GRD model is then given by

$$u(x, t; \mathcal{P}) = \int_{g \in G} v(x, t; g) d\mathcal{P}(g), \quad (2.3)$$

where  $v(x, t; g)$  is the solution to (2.1) with growth rate  $g$ ,  $G$  is the collection of admissible growth rates, and  $\mathcal{P}$  is a probability measure on  $G$ . Thus, growth in a given population is not characterized by a single growth rate function. Rather, populations are characterized by growth as embedded in the pair  $(G, \mathcal{P})$ , a set of possible growth rates and a probability distribution on this family of growth rates. This model was shown to be sufficiently rich to allow prediction of both the dispersion and bifurcation typically observed in mosquitofish (as well as other) population data [7, 8, 12, 13]. Theoretical and computational frameworks for the inverse problem of selecting a best  $\mathcal{P}$  to fit the model (2.3) to aggregate population data are developed and discussed in [4, 6, 8, 9, 14], where one endows the space of probability measures with the Prohorov metric topology [4, 17]. Specifically, one formulates the problem of finding  $\mathcal{P} \in \mathbf{P}(G)$ , (where  $\mathbf{P}(G)$  is the set of all probability distribution functions over  $G$ ), given population density data  $\hat{d}_{k,j}$ ,  $k = 1, \dots, N_t, j = 1, \dots, N_x$ , by minimizing the least squares functional

$$J(\mathcal{P}) = \sum_{k,j} \left| \hat{d}_{k,j} - \int_{g \in G} v(x_j, t_k; g) d\mathcal{P}(g) \right|^2 \quad (2.4)$$

over all  $\mathcal{P} \in \mathbf{P}(G)$ .

In summary, the GRD model (2.3) represents one approach to accounting for variability in growth rates by imposing a probability distribution on the growth rates in the SS model (2.1). Individuals in the population grow according to a deterministic growth model (2.2), but different individuals in the population may have different parameter dependent growth rates in the GRD model. The population is assumed to consist of subpopulations with individuals in the same subpopulation having the same growth rate. The growth uncertainty of individuals in the population is the result of variability in growth rates among the subpopulations. This modeling approach, which entails a stationary probabilistic structure on a family of deterministic dynamic systems, may be most applicable when the growth of individuals is assumed to be the result of genetic variability such as considered in [19].

However, a second approach that has been studied as well is based on the assumption that individual growth is a Markov diffusion stochastic process which leads to the Fokker-Planck model for shrimp population density [1, 15, 22, 29]. The growth process for each individual is stochastic, and each individual grows according to a stochastic growth model. In the Fokker-Planck model, the uncertainty in the growth of individuals is the result of the growth stochasticity of each individual. This modeling approach may be most applicable when the variability in the growth rate of individuals is believed to be the result of changes in environmental factors such as discussed in [23, 26, 30]. Theoretical arguments in [11]

demonstrate that the population density from the GRD model is the same as the population density obtained from the Fokker-Planck model when equivalent levels of variability are used in both models. Numerical results are also presented in [10] to further validate the theoretical analysis of [11]. Therefore, one can use the computationally “easier” approach to model the population of interest when appropriately chosen forms of variability can be determined. Based on these studies, we chose to use the GRD model (2.3) to incorporate uncertainty in the growth rates in the size-structured population model for the early growth of shrimp. A natural question arises immediately: how to collect data  $\hat{d}_{k,j}$  to carry out the minimization in (2.4) to determine a reasonable value for  $\mathcal{P}$ . In particular, what sampling size and sampling frequency should be used in experiments to adequately estimate  $\mathcal{P}$ ?

### 3 Experimental Design

In this section we outline an approach for determining the sampling size and the number of sampling time points one needs to obtain reliable estimates for probabilistic growth rate parameters in the GRD model. We also present some computational results demonstrating the effect of the bin size used in inverse problem calculations on parameter estimates and model predictions. We do this in the context of a parameterized family  $\mathbf{P}$  of admissible probability distributions (in this case a family of truncated normal distributions on intrinsic growth rates).

Before discussing the results, we first describe the simulated data used in these calculations. The experimental data to be used in the inverse problem calculations will be aggregate type longitudinal data (similar to the data shown in Figures 1 and 2). Individual shrimp are randomly sampled from the population at each time point; however, there is no guarantee that the same set of individuals are sampled at different time points. The simulated data that we use in the calculations in this section will also be aggregate longitudinal data.

We assume that the mortality rate  $m(x, t)$  and reproduction rate  $\beta(x, t)$  in (2.1) are both zero because we only consider the early growth dynamics of the shrimp. We also assume that the growth rate function has the form

$$\frac{dx}{dt} = g(x) = b(x + c), \tag{3.1}$$

where  $b$  represents the intrinsic growth rate of the individuals and  $c$  is a fixed constant. This growth rate function was shown to be reasonable in [11] where the average size (weight in grams) data for fifty randomly sampled shrimp was fit with the exponential function corresponding to the solution of (3.1). In order to satisfy the assumption of varying growth rates in the GRD model, we assume that the intrinsic growth rate is a random variable taking values in a compact set  $B = [\underline{b}, \bar{b}]$  with probability density function  $\varphi(b)$ . Here, we chose it to be a truncated normal (Gaussian) distribution with mean  $\mu_b$  and standard deviation  $\sigma_b$ , denoted by  $\mathcal{N}_{[\underline{b}, \bar{b}]}(\mu_b, \sigma_b^2)$ . This choice was based on previous analysis [11] that demonstrated that an assumption of a normal distribution on the intrinsic growth rates leads to a lognormal

distribution in size, which is typical of data collected on shrimp populations such as those in Figures 1 and 2. Therefore, the population density in (2.3) is given by

$$u(x, t; \mu_b, \sigma_b) = \int_{\underline{b}}^{\bar{b}} v(x, t; b) \varphi(b; \mu_b, \sigma_b) db = \int_{\underline{b}}^{\bar{b}} v(x, t; b) \frac{\frac{1}{\sigma_b} \phi\left(\frac{b-\mu_b}{\sigma_b}\right)}{\Phi\left(\frac{\bar{b}-\mu_b}{\sigma_b}\right) - \Phi\left(\frac{\underline{b}-\mu_b}{\sigma_b}\right)} db, \quad (3.2)$$

where  $\phi$  is the probability density function of the standard normal distribution,  $\Phi$  is its corresponding cumulative distribution function and  $v(x, t; b)$  is the solution of (2.1) with (3.1).

While individual shrimp are randomly sampled and weighed at each time point, the data used in the inverse problem calculations is the total number of shrimp in each size class. Let  $\Delta x$  be the length of the size class interval that we choose for each size class bin. Then the total number of population  $p(x, t; q)$  in each size class bin is approximated by

$$p(x, t; q) \approx u(x, t; q) \Delta x, \quad (3.3)$$

where  $q = (\mu_b, \sigma_b)$ , and  $u(x, t; q)$  is obtained from (3.2).

Let the sampling time points be given by  $t_k$ ,  $k = 1, 2, \dots, N_t$ . At each time point  $t_k$ , we independently draw  $N$  samples from  $\mathcal{N}_{[\underline{b}, \bar{b}]}(\mu_b, \sigma_b^2)$ ; that is, we obtain  $N$  samples of intrinsic growth rates  $b_i^k$ ,  $i = 1, 2, \dots, N$ . We also independently draw  $N$  samples of initial sizes  $x_{0,i}^k$  from a uniform distribution on the interval  $[\underline{x}_0, \bar{x}_0]$ , where  $\underline{x}_0$  and  $\bar{x}_0$  are some constants. Let  $s_i^k$  represent the size of shrimp at time  $t_k$  with intrinsic growth rate  $b_i^k$  and initial size  $x_{0,i}^k$ . Then by solving (3.1), we have

$$s_i^k = (x_{0,i}^k + c) \exp(b_i^k t_k) - c.$$

For convenience, we reorder  $\{s_i^k\}$  in increasing order and reorder  $\{b_i^k\}$  in the same order as  $\{s_i^k\}$ . To avoid introducing more notation, we continue to use the same notation; that is,

$$s_1^k \leq s_2^k \leq \dots \leq s_N^k.$$

We then group  $s_i^k$ ,  $i = 1, 2, \dots, N$  into size classes based on  $\Delta x$ . For example, if  $|s_1^k - s_2^k| < \Delta x$ , then  $s_1^k$  and  $s_2^k$  are in the same size class. Suppose that we have  $N_x^k$  size classes at time  $t_k$  after we group  $\{s_i^k\}$ . We use  $[x_j^k, x_j^k + \Delta x)$  to denote size class  $j$ ,  $j = 1, 2, \dots, N_x^k$  and  $z_j^k$  to represent

the total number of population in size class  $j$  at time  $t_k$  (that is,  $\sum_{j=1}^{N_x^k} z_j^k = N$ ,  $k = 1, 2, \dots, N_t$ ).

The center point  $x_j^k + \Delta x/2$  of size class  $j$  is used in the computations to estimate the parameters  $(\mu_b, \sigma_b)$  in order to eliminate any left bias by using  $x_j^k$  or right bias by using  $x_j^k + \Delta x$ . Then the estimate  $q^* = (\hat{\mu}_b, \hat{\sigma}_b)$  of the underlying ‘‘true’’ parameters  $q_0 = (\mu_b^0, \sigma_b^0)$  (assumed to exist) can be calculated by

$$q^* = \arg \min_{q \in Q} J(q) = \sum_{k=1}^{N_t} \sum_{j=1}^{N_x^k} |p(x_j^k + \Delta x/2, t_k; q) - z_j^k|^2,$$

where  $Q$  is some closed set in  $\mathbb{R}_+^2$ .

### 3.1 Influence of Sampling Size and Frequency on Parameter Estimation Problem

We carried out a series of inverse problem calculations in order to determine the influence of the sampling size ( $N$ ) and the number of sampling time points ( $N_t$ ) on the quality of the parameter estimation. The *sampling frequency* varied between twice a week, once a week and once every two weeks, while *sampling size*  $N$  varied between 25, 50, 75 and 100. We considered these sampling frequencies and sizes to determine minimal effort experiments to provide data sufficient to accurately estimate variability (i.e., to estimate  $q_0 = (\mu_b^0, \sigma_b^0)$ ) in the growth rates of the shrimp. A *sampling period* of six weeks was used in all of the following simulations. We also considered different  $\sigma_b^0$  values in order to determine the effect of the amount of variability in the growth rates on the *sampling size* and *sampling frequency* necessary for the experiments. The true standard deviation  $\sigma_b^0$  used to generate the simulated data was set to  $\sigma_{scale} \cdot \mu_b^0$ , where  $\sigma_{scale} = 0.1, 0.5$ , and  $0.9$ . Therefore, we had a total of 36 scenarios. The following parameter values were used in the inverse problem calculations:  $\Delta x = 0.01$ ,  $c = 0.1$ ,  $\mu_b^0 = 0.045$ ,  $\sigma_b^0 = \sigma_{scale} \cdot \mu_b^0$ ,  $\underline{b} = 0.001$ ,  $\bar{b} = 0.2$ ,  $\underline{x}_0 = 0$ ,  $\bar{x}_0 = 0.02$ . We used the following function for the initial condition  $v_0(x; g)$  :

$$v_0(x; g) = \begin{cases} \frac{N}{2\epsilon} & \text{if } x_0 - \epsilon < x < x_0 + \epsilon, \\ 0, & \text{otherwise,} \end{cases}$$

where  $\epsilon = 0.01$ . The inverse problem calculations were performed 500 times for each scenario with different sets of  $\{b_i^k\}_{i=1, k=1}^{N, N_t}$  sampled from the truncated normal distribution on  $[\underline{b}, \bar{b}]$  with mean  $\mu_b^0$  and standard deviation  $\sigma_b^0$  and  $\{x_{0,i}^k\}_{i=1, k=1}^{N, N_t}$  sampled from a uniform distribution on  $[\underline{x}_0, \bar{x}_0]$ . The estimated value for  $q_0$  in the  $j$ th ( $j = 1, 2, \dots, 500$ ) inverse problem of the  $l$ th ( $l = 1, 2, \dots, 36$ ) scenario is denoted as  $q^{*j,l} = (\hat{\mu}_b^{j,l}, \hat{\sigma}_b^{j,l})$ .

Figure 3 displays the results obtained when  $\sigma_{scale} = 0.1$ . The upper two plots show the average relative errors of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ , while the 95% confidence bounds for  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  are in the lower two plots of Figure 3. The average relative errors  $RE(\hat{\mu}_b)^l$  and  $RE(\hat{\sigma}_b)^l$  of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ , respectively, in each scenario is given by

$$RE(\hat{\mu}_b)^l = \frac{1}{500} \sum_{j=1}^{500} \frac{|\hat{\mu}_b^{j,l} - \mu_b^0|}{\mu_b^0} \quad \text{and} \quad RE(\hat{\sigma}_b)^l = \frac{1}{500} \sum_{j=1}^{500} \frac{|\hat{\sigma}_b^{j,l} - \sigma_b^0|}{\sigma_b^0}, \quad \text{for } l = 1, 2, \dots, 36.$$

We obtain the 95% confidence bounds for  $\hat{\mu}_b^l$  and  $\hat{\sigma}_b^l$  for each scenario by computing the value where 2.5% of the estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  are below and above based on the results of the 500 inverse problem calculations. While the *sampling size*  $N$  and *sampling frequency* do not have an effect on the estimation of  $\mu_b$  in this example, we observe that there is an effect on the estimation of  $\sigma_b$ . The average relative error of  $\hat{\sigma}_b$  decreases as we increase either  $N$  or  $N_t$ . Furthermore, the reduction in the average relative error resulting from an increase in the *sampling size*  $N$  from 25 to 50 is greater than the reduction caused by an increase in  $N$  from 50 to 75 and 75 to 100. These results suggest that a *sampling size* of 25 is not sufficiently large to obtain reliable estimates of the variability in the growth rates.



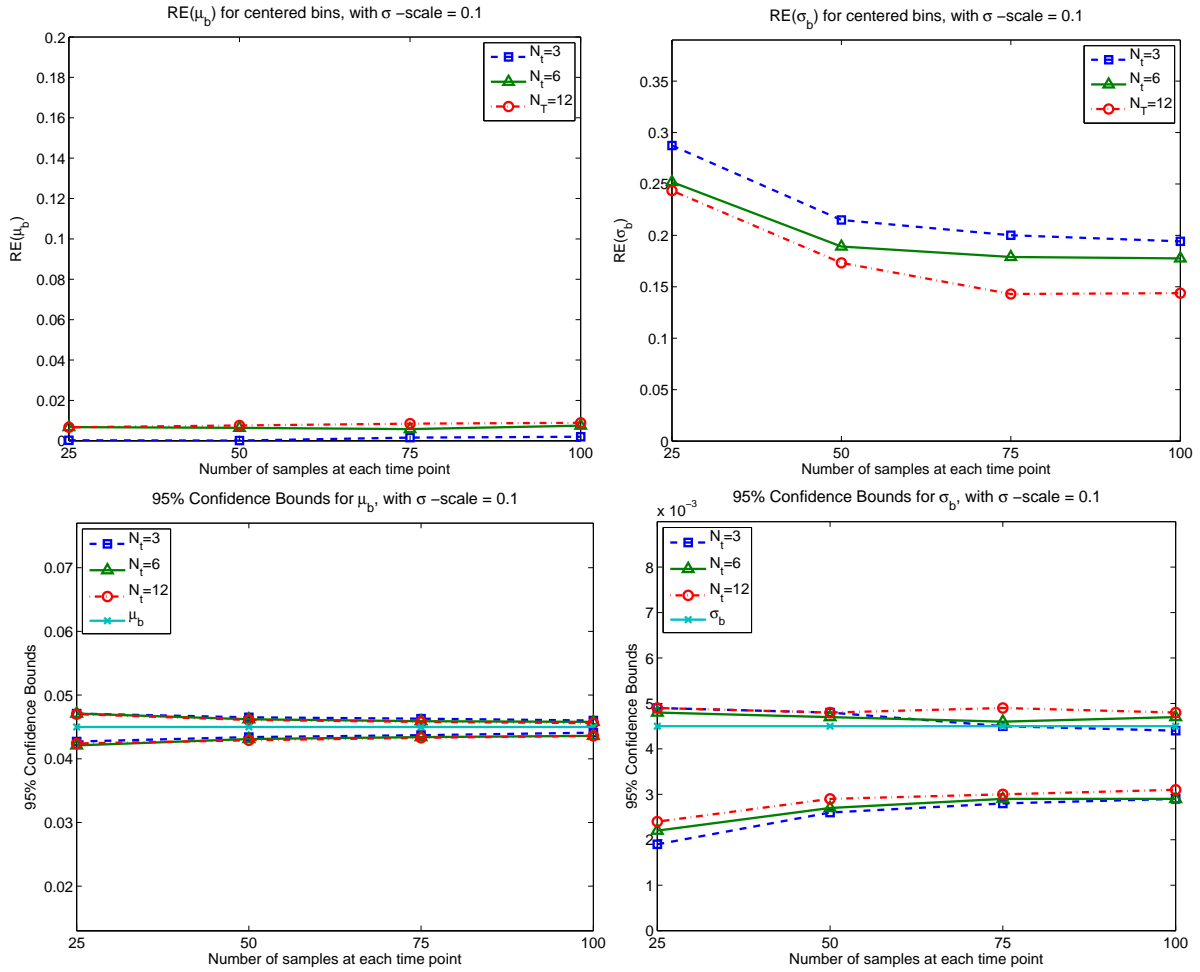


Figure 3: (Upper): Relative errors of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ ; (Lower): 95% confidence bounds for the estimates corresponding to  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  with  $\sigma_{scale} = 0.1$ .

The lower plots in Figure 3 depicting the 95% confidence bounds for  $\hat{\mu}_b$  also show that for this example with  $\sigma_{scale} = 0.1$  the *sampling size* and *frequency* have minimal effect on the estimates  $\hat{\mu}_b$  based on the very tight bounds around the “true” mean  $\mu_b^0$ . However, we observe a decrease in the width of the confidence bounds for  $\hat{\sigma}_b$  as  $N$  is increased from 25 to 50 and essentially no change in the width as  $N$  is increased from 50 to 75 and 75 to 100. This also suggests that  $N = 25$  is not sufficient to obtain a reliable estimate of  $\sigma_b$ .

A  $\sigma_{scale}$  value of 0.5 was used in the simulations that produced the results in Figure 4. We note in this case that the estimates of  $\mu_b$  and  $\sigma_b$  are both affected by the *sampling size* and *frequency*. The average relative error of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  both decrease as  $N$  is increased. The decrease in the average relative error is very significant as  $N$  is increased from 25 to 50 for both parameters; however, the reduction in the average relative error for both parameters is not as dramatic as  $N$  is increased from 50 to 75 and from 75 to 100. We would again infer from these results that a sampling size of at least 50 is desirable in order to obtain reliable estimates of  $(\mu_b, \sigma_b)$ .

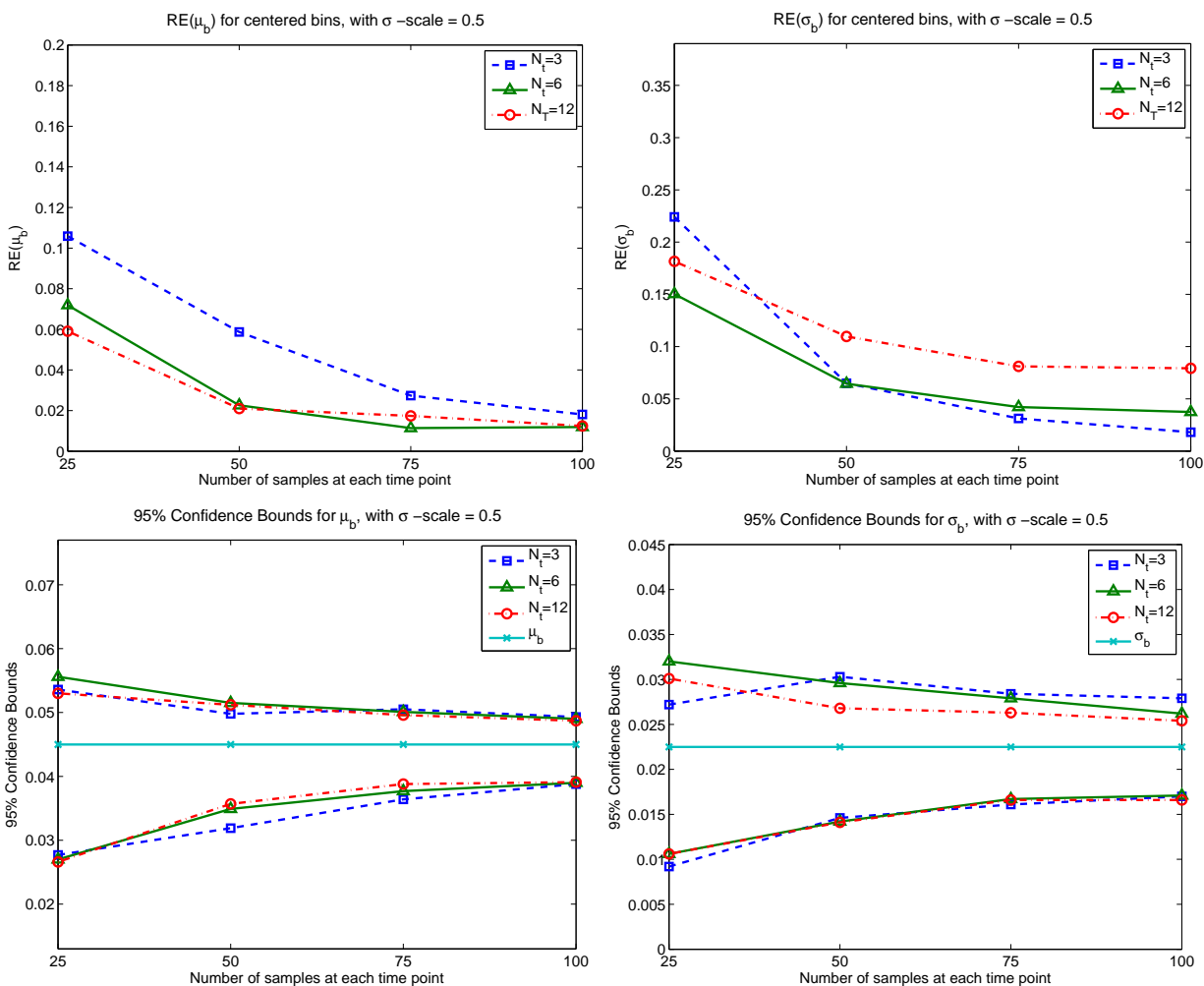


Figure 4: (Upper): Relative errors of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ ; (Lower): 95% confidence bounds for the estimates corresponding to  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  with  $\sigma_{scale} = 0.5$ .

The resulting 95% confidence bounds for  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  when  $\sigma_{scale} = 0.5$  are shown in the lower plots of Figure 4. As already noted, the estimation of both parameters depends on  $N$  and  $N_t$ . The decrease in the width of the confidence bounds for  $\hat{\mu}_b$  is more significant as  $N$  is increased from 25 to 50 in comparison to the decrease in the width when  $N$  is increased from 50 to 75 and 75 to 100. We also observe the same type of behavior in the width of the confidence bounds for  $\hat{\sigma}_b$ . The reduction in the width of the confidence bounds for  $\hat{\sigma}_b$  is much smaller as  $N$  is increased from 50 to 75 and 75 to 100 versus the decrease in width as  $N$  is increased from 25 to 50. While a *sampling size*  $N$  of 25 appears to be too small, a *sampling size* of 50 seems to be sufficient to obtain reasonably accurate estimates of the variability in the growth rates.

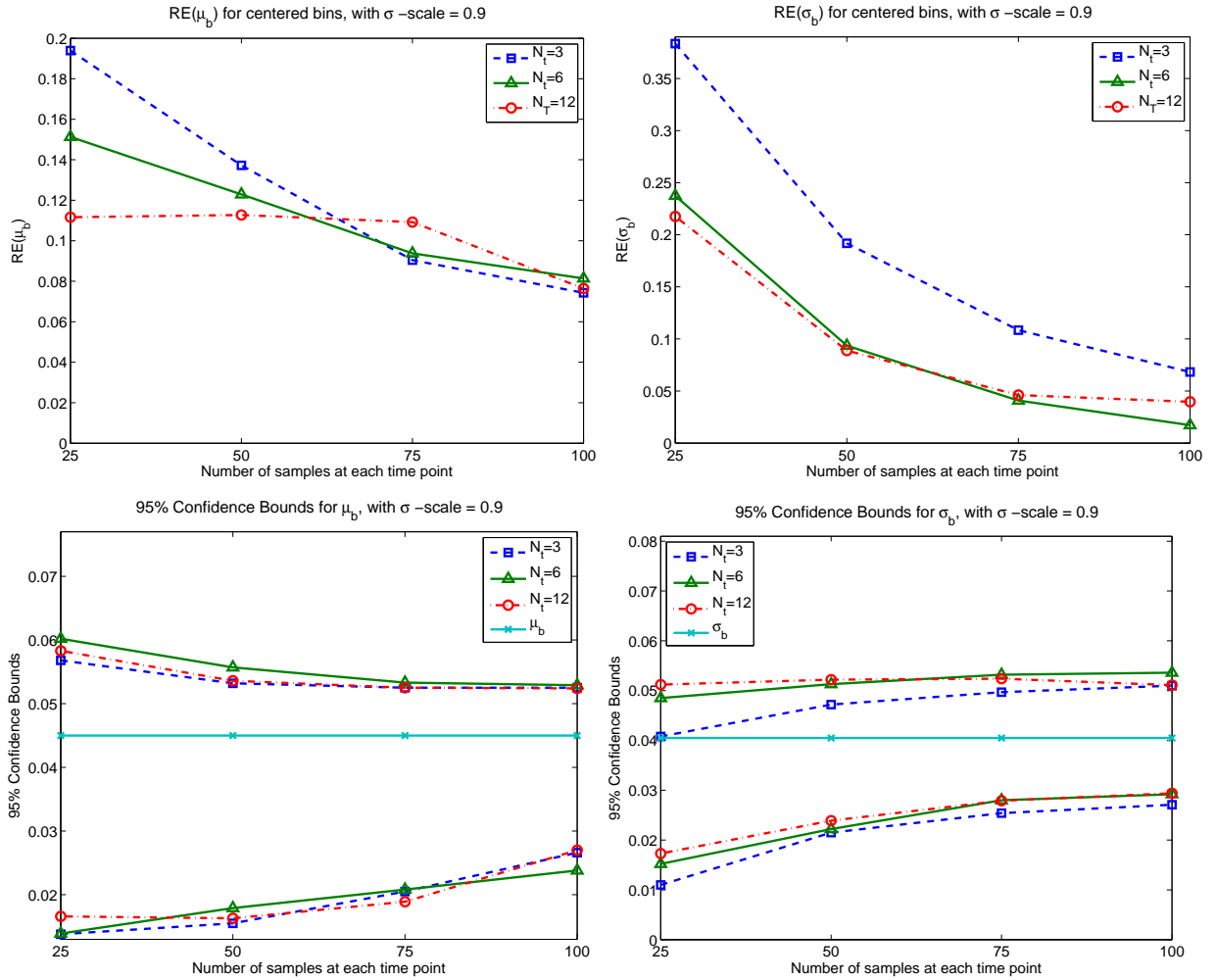


Figure 5: (Upper): Relative errors of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ ; (Lower): 95% confidence bounds for the estimates corresponding to  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  with  $\sigma_{scale} = 0.9$ .

The final sets of results in Figure 5 were computed with  $\sigma_{scale} = 0.9$ . The average relative errors for  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  depicted in the upper plots in Figure 5 reveal that the reliability of both parameters is influenced by the *sampling size* and *frequency*. As the previous results suggested, a *sampling size* of 25 is not adequate for the parameter estimation problem based

on the magnitude of the average relative error for  $N = 25$  in comparison to the average relative errors obtained for  $N = 50, 75,$  and  $100$ . We also note that the decrease in the average relative error of  $\hat{\sigma}_b$  is not as large when  $N$  is increased from 50 up to 100 in comparison to a much greater decrease when  $N$  is increased from 25 to 50. These results suggest that there is not a significant improvement in the reliability of the parameter estimates when a *sampling size* larger than 50 is used. We also observe from the lower plots in Figure 5 that the effect of  $N$  on the width of the 95% confidence bounds of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  decreases as  $N_t$  is increased. As in the previous two cases considered, the width of the confidence bounds for both parameters decreases the most when  $N$  is increased from 25 to 50 which also appears to suggest that reliable estimates of  $(\mu_b, \sigma_b)$  can be obtained with a *sampling size*  $N$  of 50.

Based on these simulations, we conclude that, not surprisingly, the *sampling size*  $N$  does have an effect on the reliability of the results of the parameter estimation problem. When the standard deviation  $\sigma_b^0$  is small relative to the mean  $\mu_b^0$ , the effect on the accuracy of the estimates  $\hat{\mu}_b$  is minimal; however, this effect is more noticeable as the magnitude of  $\sigma_b^0$  is increased. The influence of the *sampling size*  $N$  is much more evident in the estimation of  $\sigma_b$  for each of the values considered here. We also conclude that the results when sampling once a week are comparable to those obtained when sampling twice a week. The most desirable experiment involved using a *sampling size*  $N$  of 100 once a week; however, there appears to be little loss in accuracy in estimating the variability in growth rates if one uses a sample size  $N = 50$ . Finally, our computations suggest that experiments with a *sampling size* of only  $N = 25$  would not be adequate even when one increases the *sampling frequency* to twice a week.

### 3.2 Effect of Sampling Size and Bin Size on Model Predictions and Parameter Estimates

We also carried out numerical simulations with simulated data to investigate the effect of the *bin size*  $\Delta x$  in relation to the *sampling size*  $N$  on the parameter estimation results as well as the model predictions. In this section we also compare the results from computations with a fixed bin size versus variable bin sizes (fixed number of bins  $N_b$ ) for a fixed *sampling size* in the estimation problem. Since the “true” growth rate distribution of shrimp observed in experiments is unknown, the accuracy of the parameter estimates is determined by how well the model fits the data. The discussion of the results for the inverse problem with simulated data in this section will emphasize the importance in choosing the appropriate *bin size*  $\Delta x$  with respect to the *sampling size*  $N$  in the inverse problem calculations.

In the first set of results we used a fixed  $\Delta x$  with two different  $N$  values to illustrate the significance of using the proper bin size relative to the sample size for accurate model predictions. Using simulated data generated with a truncated normal distribution with mean  $\mu_b^0 = 0.045$  and standard deviation  $\sigma_b^0 = 0.5 \cdot \mu_b^0 = 0.0225$ , we obtained the following results from the inverse problem with a fixed *bin size*  $\Delta x$  of 0.01. Because of our earlier findings discussed in the last section, we used a *sampling size*  $N$  of 50 with *sampling frequency* of once

a week for six weeks. Using our inverse problem techniques, we computed good parameter estimates  $(\hat{\mu}_b, \hat{\sigma}_b)$  with small relative errors (see Table 1) which suggests that the model does a good job of fitting the data. However, the model does not provide good fits to the data as time progresses based on the plots shown in Figure 6. The model appears to do a good job of predicting the population data during the first two weeks but fails to predict the magnitudes seen in the data during the final four weeks. It is clear from Figure 6 that the model performs reasonably qualitatively; however, as a result of the quantitative comparisons of the model predictions to the data, the model fits do not support any optimism based on the numerical results in Table 1.

N	$\hat{\mu}_b$	$RE(\hat{\mu}_b)$	$\hat{\sigma}_b$	$RE(\hat{\sigma}_b)$
50	0.04817	0.07044	0.02400	0.06667

Table 1: Estimated mean and standard deviation  $(\hat{\mu}_b, \hat{\sigma}_b)$  and relative errors for inverse problem with simulated data with  $N = 50$  and  $b \sim \mathcal{N}_{[0.001, 0.2]}(0.0450, 0.0225^2)$  and  $\Delta x = 0.01$ .

We then generated simulated data with a *sampling size* of 5000 and a truncated normal distribution on the intrinsic growth rates with mean  $\mu_b^0 = 0.045$  and standard deviation  $\sigma_b^0 = 0.0225$ . The estimated parameters along with the corresponding relative errors that were computed with the fixed bin size  $\Delta x = 0.01$  are in Table 2.

N	$\hat{\mu}_b$	$RE(\hat{\mu}_b)$	$\hat{\sigma}_b$	$RE(\hat{\sigma}_b)$
5000	0.04495	0.00111	0.02253	0.00133

Table 2: Estimated mean and standard deviation  $(\hat{\mu}_b, \hat{\sigma}_b)$  and relative errors for inverse problem with simulated data with  $N = 5000$  and  $b \sim \mathcal{N}_{[0.001, 0.2]}(0.0450, 0.0225^2)$  and  $\Delta x = 0.01$ .

Based on the small relative errors, we also conclude in this example that we have obtained good estimates of the true parameters  $(\mu_b^0, \sigma_b^0)$ . Moreover, we assume that the model predictions would fit the data very well. Figure 7 presents plots of the model predictions versus the simulated data for six weeks with a *sampling size* of 5000. We see that the model does a very good job both qualitatively and quantitatively. Unlike the previous example with  $N = 50$ , the model provides good fits to the entire data set when  $N = 5000$ .

These results demonstrate the effect of the bin size relative to the *sampling size*  $N$  on the accuracy of the model predictions. While the parameter estimates obtained for  $(\mu_b^0, \sigma_b^0)$  with both sampling sizes were good, the model predictions with a sample *size* of 50 were not as accurate as those obtained with a sample *size* of 5000. As noted earlier, the accuracy of  $(\hat{\mu}_b, \hat{\sigma}_b)$  will depend on the model fits to data when experimental data is used since the “true” underlying distribution of the growth rates is unknown and relative errors cannot be computed. Therefore, using an appropriately chosen *bin size*  $\Delta x$  is critical in the parameter estimation problem with experimental data.

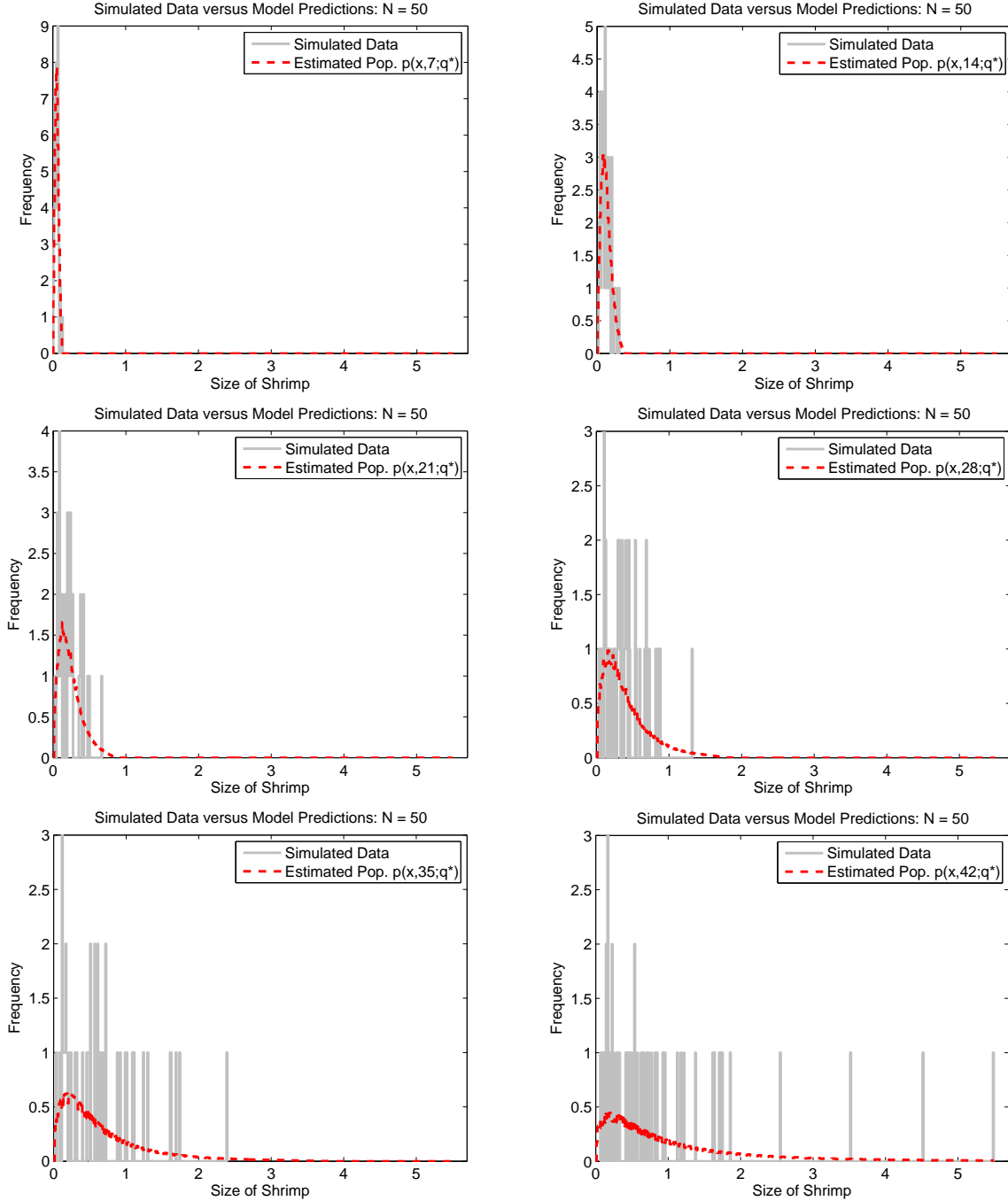


Figure 6: Simulated population data, snapshot  $z_j^k$ , and model predicted population  $p(x, t; q^*)$  with optimal parameter  $q^*$  for  $t = 7, 14, 21, 28, 35, 42$  days with sample size  $N = 50$ .

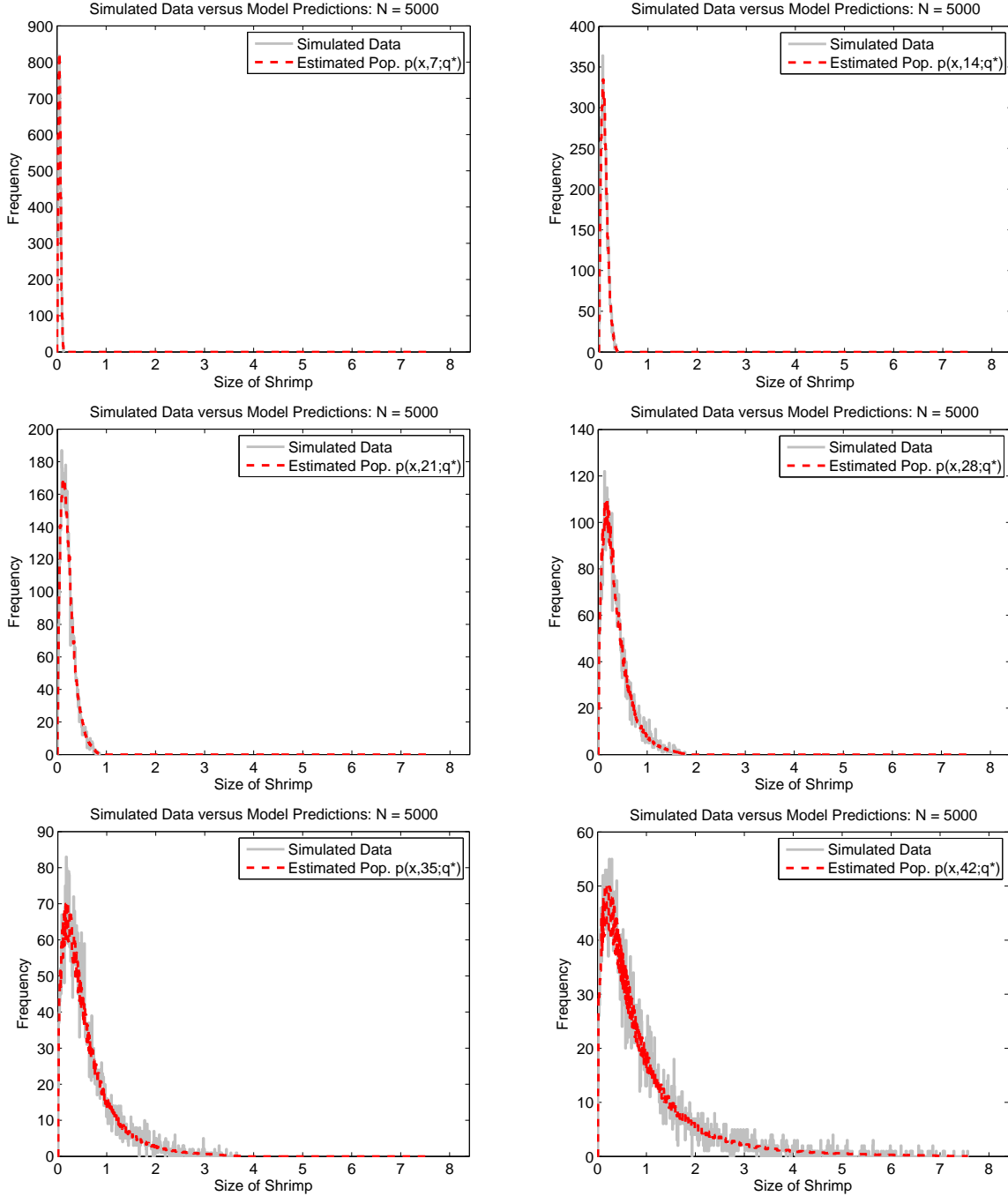


Figure 7: Simulated population data, snapshot  $z_j^k$ , and model predicted population  $p(x, t; q^*)$  with optimal parameter  $q^*$  for  $t = 7, 14, 21, 28, 35, 42$  days with sample size  $N = 5000$ .

To further investigate the effect of  $\Delta x$  on the accuracy of the model predictions and parameter estimates, we considered using a fixed *bin size*  $\Delta x$  versus a fixed number  $N_b$  of bins (variable bin size) in the parameter estimation problem with simulated data for a fixed *sampling size*. A *sampling size*  $N$  of 50 with a *sampling frequency* of once a week for six weeks was used in the following results. The simulated data was generated with a truncated normal distribution with mean  $\mu_b^0 = 0.045$  and standard deviation  $\sigma_b^0 = 0.0225$ . Using the same data set, we computed results with a fixed bin size  $\Delta x = 0.01$  and with a fixed number  $N_b = 10$  of bins. The bin size for each time point when a fixed number of bins is used is given by

$$\Delta x(t_k) = \frac{\bar{x}(t_k) - \underline{x}(t_k)}{N_b},$$

where  $\bar{x}(t_k)$  is the largest size sampled at time  $t_k$  and  $\underline{x}(t_k)$  is the smallest size sampled at time  $t_k$ . The parameter estimates  $(\hat{\mu}_b, \hat{\sigma}_b)$  along with the relative errors for both scenarios are given in Table 3. While we obtain comparable estimates of the mean  $\mu_b^0$ , we note that the estimated standard deviation  $\hat{\sigma}_b$  is better when using a fixed number of bins versus a fixed bin size. We observe from the model predictions versus the data in Figure 8 that as time evolves the model is not able to predict the peaks present in the data when the fixed bin size  $\Delta x = 0.01$  is used. However, the model predictions obtained with the fixed number  $N_b = 10$  of bins (varying bin size) are more accurate as seen in Figure 9. The model does a much better job of fitting the data when a fixed number of bins are used versus a fixed bin size. We infer from the results shown in Figure 8 that the fixed bin size  $\Delta x = 0.01$  is too small; this leads to the increased number of size classes with a frequency equal to one seen in the simulated data. This is not the case with the fixed number  $N_b = 10$  of bins.

	$\hat{\mu}_b$	$RE(\hat{\mu}_b)$	$\hat{\sigma}_b$	$RE(\hat{\sigma}_b)$
$\Delta x = 0.01$	0.04471	0.00644	0.01965	0.12667
$N_b = 10$	0.04485	0.00333	0.02146	0.04622

Table 3: Estimated mean and standard deviation  $(\hat{\mu}_b, \hat{\sigma}_b)$  and relative errors for inverse problem with simulated data with  $b \sim \mathcal{N}_{[0.001, 0.2]}(0.045, 0.0225^2)$ .

Similar results were obtained when we repeated the inverse problem calculations 500 times. Using a fixed bin size that is too small results in model predictions that are not very accurate during the early growth process of the shrimp when the *sampling size*  $N$  is relatively small. Although the plots in Figure 9 demonstrate the benefits of fixing the number of bins, one must still be careful when selecting the number of bins to use in computations because choosing a value that is too small or too large can bias the data as well and affect the model predictions. Furthermore, we also found that a larger fixed *bin size*  $\Delta x$  can result in better model predictions. The results shown here demonstrate the importance in using the appropriate number  $N_b$  of bins or *bin size*  $\Delta x$  in ratio to the *sampling size*  $N$  to obtain good model fits and parameter estimates in inverse problem calculations.

The simulated inverse problem investigations discussed in this and the previous section are not only helpful in designing experiments by suggesting sample size and frequency, but they



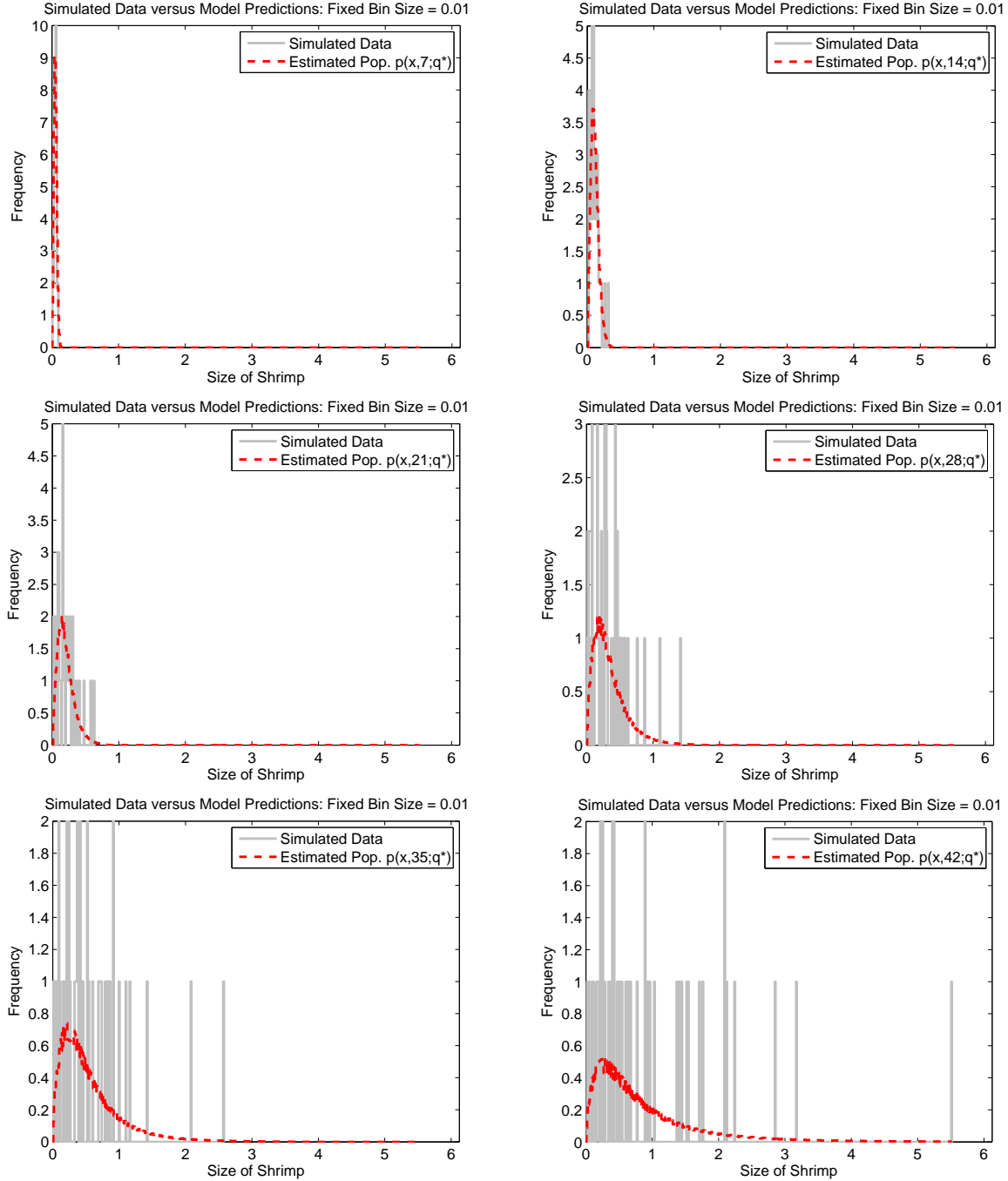


Figure 8: Simulated population data, snapshot  $z_j^k$ , and model predicted population  $p(x, t; q^*)$  with optimal parameter  $q^*$  for  $t = 7, 14, 21, 28, 35, 42$  days with fixed bin size  $\Delta x = 0.01$ .

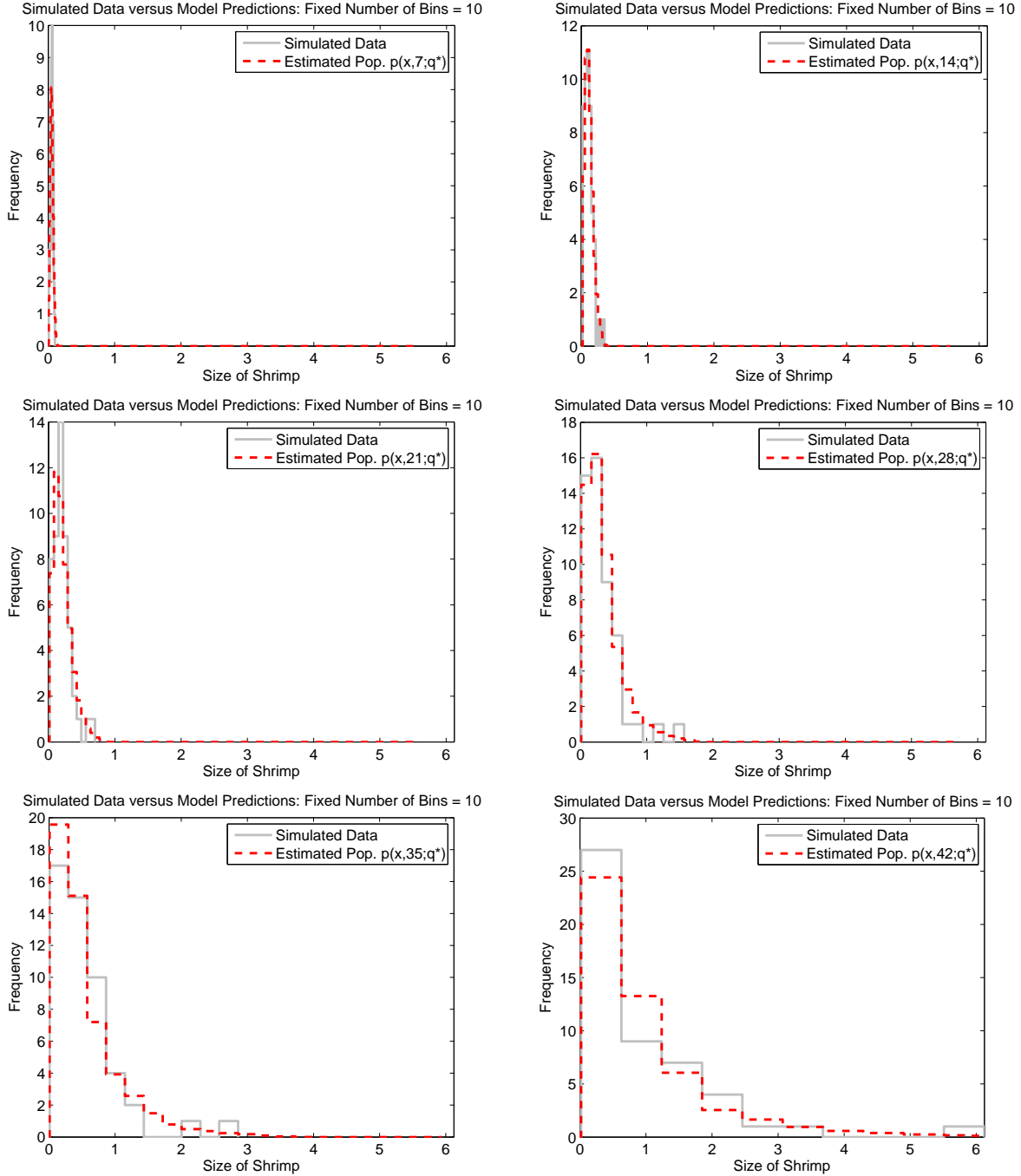


Figure 9: Simulated population data, snapshot  $z_j^k$ , and model predicted population  $p(x, t; q^*)$  with optimal parameter  $q^*$  for  $t = 7, 14, 21, 28, 35, 42$  days with fixed number of bins  $N_b = 10$ .

also suggest the importance of careful inverse problem formulation once experimental data is obtained. In particular, the manner in which a given data set is used in the inverse problem formulation can be very important in determining the quality of estimates obtained.

## 4 Computational Results for Parameter Estimation Problem with Experimental Data

Based on the design calculations discussed in Section 3, experiments were carried out in controlled environment tanks at ABN. In this section, we present the results of the inverse problem for the estimation of the growth rate distribution of a size-structured shrimp population using this data. Data were collected during the early growth of shrimp initially in the post-larval stage. Shrimp were randomly sampled from a fixed population growing under relatively constant tank conditions (density, temperature, feeding, etc.) and fifty shrimp were measured weekly from December 4 through December 31, 2007, resulting in a data set with five longitudinal size distribution observations. As noted earlier, we must compare the experimental data to the estimated population data generated by using the optimal parameter  $q^*$  in (3.3) since the actual growth rate distribution of the shrimp observed is unknown. The following results were obtained via computations carried out in MATLAB with bin size and intrinsic growth bounds  $\Delta x = 0.02$ ,  $\underline{b} = 10^{-6}$ ,  $\bar{b} = 10^{-1}$  in the context of truncated normal distributions as described in the simulation studies of Section 3.

Before presenting the results, we briefly discuss how we obtain confidence intervals for the parameter estimates based on the asymptotic distribution of ordinary least squares (OLS) estimators for finite dimensional parameters [20, 21, 24, 31]. We assume the following statistical model describes the observations

$$Y_j = Y(\tau_j) = f(\tau_j; \theta_0) + \epsilon_j, \quad j = 1, \dots, n, \quad (4.1)$$

where  $f(\tau_j; \theta_0)$  represents the observed mathematical model response (3.3),  $\tau_j$  represents  $(x, t)$  pairs, and  $\theta_0 \in \Theta \subset \mathbb{R}^2$  represents the “true” parameter value (unknown but assumed to exist in the usual statistical framework we employ here). We note that for our examples  $\theta_0 \approx \theta = q = (\mu_b, \sigma_b)$ . The  $\epsilon_j$  are also assumed to be independent identically distributed (*i.i.d.*) random variables with  $E[\epsilon_j] = 0$  and  $Var[\epsilon_j] = \sigma_0^2$ , where  $\sigma_0^2 > 0$  is unknown. We note that the observations  $Y_j$  are also random variables with mean  $E[Y_j] = f(\tau_j; \theta_0)$  and variance  $\sigma_0^2$ . The OLS estimator  $\theta_{OLS}(Y)$  used in the inverse problem for the estimation of  $\theta$  is given by

$$\theta_{OLS}(Y) \equiv \arg \min_{\theta \in \Theta} \sum_{j=1}^n (Y_j - f(\tau_j; \theta))^2. \quad (4.2)$$

Because the observations  $Y_j$  are random variables, then the OLS estimator  $\theta_{OLS}(Y)$  is also a random variable, and as  $n \rightarrow \infty$ , the sampling distribution for  $\theta_{OLS}(Y)$  can be approximated by a multivariate normal distribution with mean  $E[\theta_{OLS}(Y)] \approx \theta_0$  and covariance

$Var[\theta_{OLS}(Y)] \approx \Sigma_0^n \approx \sigma_0^2[\mathcal{X}^T(\theta_0)\mathcal{X}(\theta_0)]^{-1}$ . The  $n \times 2$  sensitivity matrix  $\mathcal{X}(\theta)$  has entries

$$\mathcal{X}_{jk}(\theta) = \frac{\partial f(\tau_j; \theta)}{\partial \theta_k}$$

where  $\theta = q = (\mu_b, \sigma_b)$ . Since  $\theta_0$  and  $\sigma_0^2$  are generally unknown, we must use estimates for these parameters in order to approximate the covariance matrix  $\Sigma_0^n$ . The estimate for  $\theta_0$  is determined by using a particular realization or data set  $\{y_j\}$  to find the OLS estimates  $\hat{\theta}$  that minimize

$$J(\theta) = \sum_{j=1}^n (y_j - f(\tau_j; \theta))^2.$$

The estimate for  $\sigma_0^2$  is given by

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^n (y_j - f(\tau_j; \hat{\theta}))^2.$$

Using the estimates  $\hat{\theta}$  and  $\hat{\sigma}^2$ , we can find an estimate of the covariance matrix  $\Sigma_0^n$  by computing  $\Sigma_0^n \approx \hat{\Sigma} = \hat{\sigma}^2[\mathcal{X}^T(\hat{\theta})\mathcal{X}(\hat{\theta})]^{-1}$ . The endpoints of the confidence intervals based on the standard errors are

$$\hat{\theta}_k \pm t_{1-\alpha/2} SE(\hat{\theta}_k), \quad k = 1, 2,$$

where the standard errors are given by

$$SE(\hat{\theta}_k) = \sqrt{\hat{\Sigma}_{kk}}, \quad k = 1, 2.$$

Here  $t_{1-\alpha/2}$  is computed from the Student's t-distribution and depends on the level of significance chosen [18]. We can use the confidence intervals to quantify the uncertainty in the estimates of the finite-dimensional parameter  $\theta$  obtained from the OLS estimator for a given data set.

For the application considered here, we note that we are able to compute the entries of the sensitivity matrix explicitly. Since

$$f(\tau_j; \theta_0) \approx p(\tau_j; \theta),$$

the entries in the sensitivity matrix  $\mathcal{X}(\theta)$  are given by

$$\begin{aligned} \mathcal{X}_{jk}(\theta) &= \frac{\partial f(\tau_j; \theta)}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} p(\tau_j; \theta) \\ &= \Delta x \int_{\underline{b}}^{\bar{b}} v(\tau_j; b) \frac{\partial \varphi(b; \theta)}{\partial \theta_k} db, \end{aligned}$$

where  $\theta = q = (\mu_b, \sigma_b)$  and  $\varphi(b; \theta)$  is the probability density function (pdf) for the truncated normal distribution. We chose to use  $\alpha = 0.05$  which corresponds to a significance level of 95% and a  $t_{1-\alpha/2}$ -value of approximately 1.96 when  $n$  is greater than or equal to 30.

## 4.1 Results with Complete December Data Set

In the first set of results we considered the entire data set collected in December, and we began by fitting the exponential function

$$\bar{x}(t) = a \exp(bt) - c \quad (4.3)$$

to the data  $\{(t_k, \bar{x}_d^k)\}$  where  $a, b$ , and  $c$  are all unknown parameters and  $\bar{x}_d^k$  is the average size of shrimp observed at time  $t_k$ . This exponential function was determined by solving the deterministic growth rate function of the form

$$\frac{dx}{dt} = b(x + c), \quad (4.4)$$

where  $b$  and  $c$  are assumed to be fixed constants. The optimal parameters corresponding to the fit with (4.3) shown in Figure 10 are  $\hat{a} = 0.049, \hat{b} = 0.043$  and  $\hat{c} = 0.036$ .

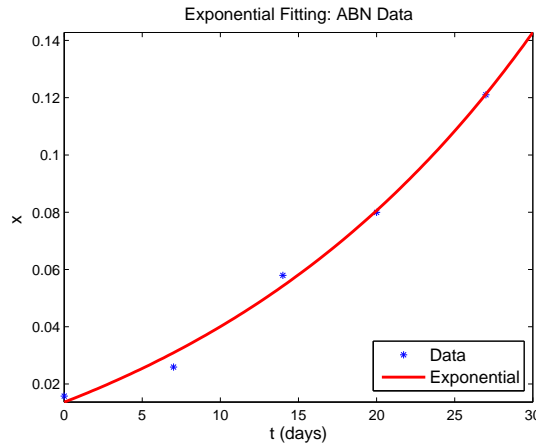


Figure 10: Exponential fit of (4.3) to complete ABN average size shrimp data set with  $\frac{d\bar{x}}{dt} = 0.043(\bar{x} + 0.036)$ .

We then used the optimal value  $\hat{c}$  in our inverse problem calculations for the estimation of  $(\mu_b, \sigma_b)$  with the entire December 2007 data. The data from December 4, 2007 (see Figure 11) is interpolated and used as an approximation for the initial size density  $v_0(x; g)$ . Therefore, it cannot be used in the estimation of the growth rate distribution, and we are left with four longitudinal data records to use in the inverse problem.

The estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  along with the confidence intervals for each of these parameters are given in Table 4. These results have an optimal cost  $J(q^*)$  of 574.8315, and the estimated variance of the statistical model is  $\hat{\sigma}^2 = 17.4191$ . We observe from the plots in Figure 11 that the model predictions do *not* fit the data well. The model does a decent job of fitting the data on December 11 but is shifted to the left of the observations for the final three time points. Moreover, the estimates  $(\hat{\mu}_b, \hat{\sigma}_b)$  are not very reliable based on the magnitude of the associated confidence intervals relative to the estimated parameter values.

$\hat{\mu}_b$	$0.0010 \pm 0.0535$	$\hat{\sigma}_b$	$0.0324 \pm 0.0313$
---------------	---------------------	------------------	---------------------

Table 4: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with complete December data set using  $\hat{c} = 0.036$  from exponential fit of (4.3) to average size shrimp data.

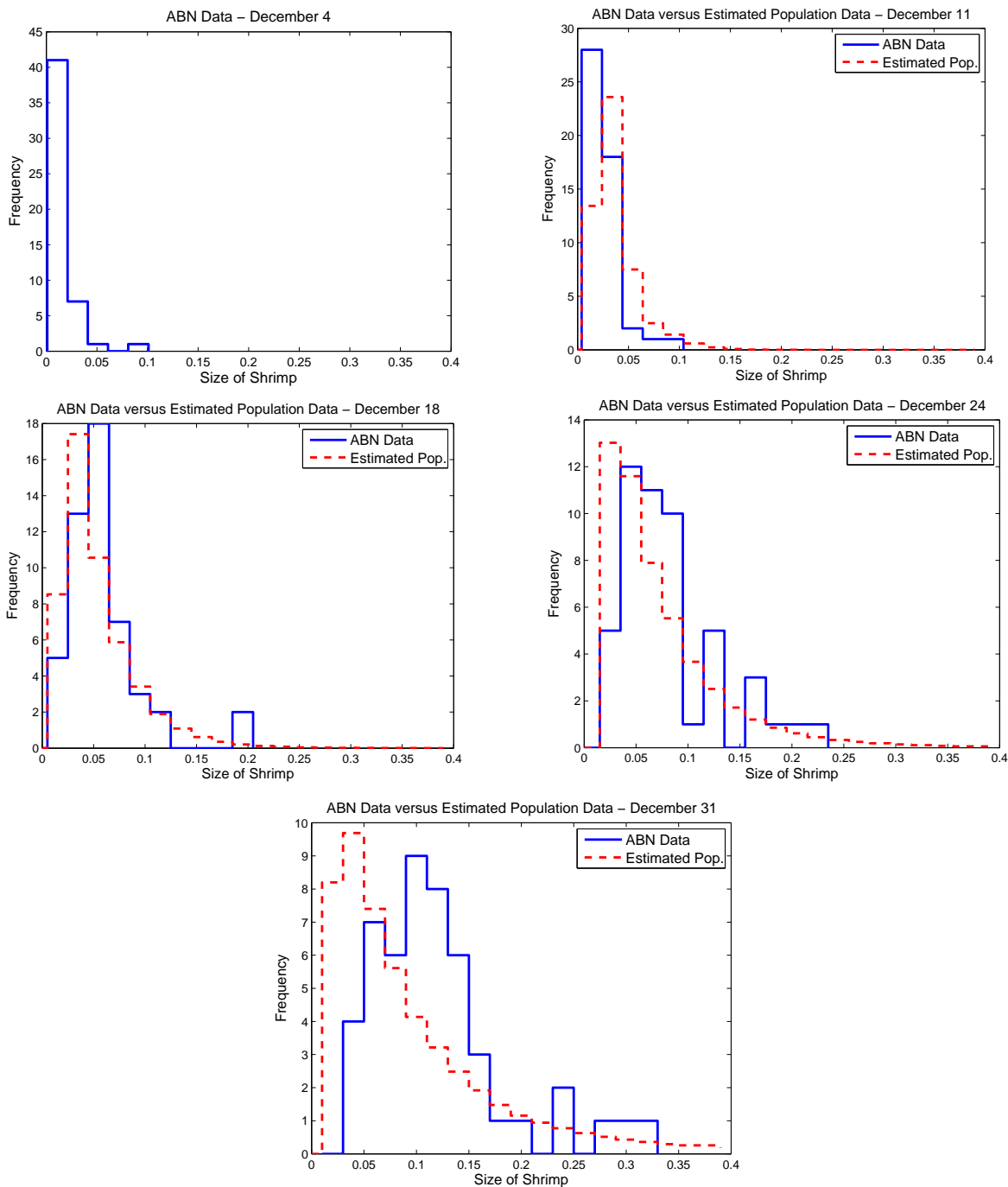


Figure 11: Longitudinal data on shrimp weight (in grams) from ABN on December 4, 11, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0010, 0.0324)$  and  $\hat{c} = 0.036$  from exponential fit of (4.3) to average size shrimp data.

In [11] the authors outline and compare the two previously mentioned approaches (a probabilistic formulation corresponding to the Growth Rate Distribution model versus a stochastic formulation corresponding to the Fokker-Planck model) for modeling variability in the growth process of size-structured populations. As established in [11], the population density from these two approaches is the same when comparable levels of uncertainty are used in both formulations. Hence, we chose to use the computationally “easier” GRD model (2.3) for our purposes. The shortcomings of the deterministic growth model (4.4) in modeling populations exhibiting a great deal of variability was also discussed in [11] with several different approaches for including uncertainty in the growth process also being described. One such approach (the one taken in this paper) is based on considering the probabilistic growth model

$$\frac{dx(t; b)}{dt} = b(x(t; b) + c), \quad b \in B, \quad (4.5)$$

where a distribution is placed on the intrinsic growth rates taking values in  $B$  and  $c$  is assumed to be a fixed affine growth term. If a normal distribution with mean  $\mu_b$  and variance  $\sigma_b^2$  is used to describe the distribution of the intrinsic growth rates, then the exponential function corresponding to probabilistic growth model (4.5) is given by

$$x(t; b) = (x(0; b) + c) \exp(bt) - c, \quad (4.6)$$

where  $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$ . Moreover, it was shown that the probabilistic growth model results in the following exponential function for the mean of the stochastic process for the size of the individuals

$$\bar{x}(t) = (\bar{x}(0) + c) \exp\{\mu_b t + \frac{1}{2} \sigma_b^2 t^2\} - c \quad (4.7)$$

which corresponds to the time-varying mean growth dynamics

$$\frac{d\bar{x}(t)}{dt} = (\mu_b + \sigma_b^2 t)(\bar{x}(t) + c). \quad (4.8)$$

As an alternative to the deterministic form (4.4), we also chose to fit the average size shrimp data  $\{(t_k, \bar{x}_d^k)\}$  with (4.7) where  $\mu_b, \sigma_b$ , and  $c$  are the parameters to be estimated and  $\bar{x}(0)$  is the average size shrimp data for the initial start time. The fit using the time varying intrinsic growth rate, shown in Figure 12, was generated with the estimated parameters  $\hat{\mu}_b = 0.045$ ,  $\hat{\sigma}_b = 0.003$  and  $\hat{c} = 0.029$ .

The data from December 4 was again used as an approximation for the initial size density  $v_0(x; g)$  in these computations. We used the estimated values of  $\hat{\mu}_b$  and  $\hat{\sigma}_b$  from the exponential fit to the average size shrimp data as initial starting points for the optimization calculations to determine the growth rate distribution of the shrimp population. The optimal value of  $c$  was also used in our inverse problem calculations for the growth rate distribution parameters. The optimal cost corresponding to the optimized parameters in Table 5 is 563.9064. We computed a value of 17.0881 for the estimate of the variance  $\sigma_0^2$  of the statistical model. The results in this case are very similar to those in the previous case as can be seen by the model fits to data shown in Figure 13. We observe the same behavior in the model predictions versus the observations in the final three time points with the model

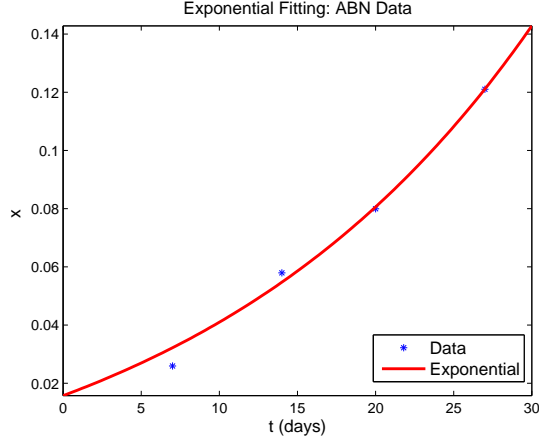


Figure 12: Exponential fit of (4.7) to complete ABN average size shrimp data with  $\frac{d\bar{x}}{dt} = (0.045 + 0.003^2t)(\bar{x} + 0.029)$ .

predictions being shifted to the left of the data. The width of the confidence intervals for the estimates  $(\hat{\mu}_b, \hat{\sigma}_b)$  also demonstrate how unreliable these parameter values are as well.

$\hat{\mu}_b$	$0.0018 \pm 0.0627$	$\hat{\sigma}_b$	$0.0363 \pm 0.0377$
---------------	---------------------	------------------	---------------------

Table 5: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with complete December data set using  $\hat{c} = 0.029$  from exponential fit of (4.7) to average size shrimp data.

It is clear from the results in this section that the model does not provide a good fit to the entire data set from December. While the model predictions do exhibit the lognormal behavior characteristic of the observations, the model does a poor job of fitting the data quantitatively. The exponential fits to the average size shrimp data in Figures 10 and 12 suggest that the problem in fitting the data could potentially be a result of the data sampled on December 11. The average size shrimp data point for December 11 lies the farthest away from the exponential curves produced by (4.3) and (4.7). Therefore, we reconsidered the inverse problem for the estimation of the growth rate distribution for the shrimp population with the data from December 4, 18, 24 and 31 in an effort to obtain better parameter estimates and model fits to the data.



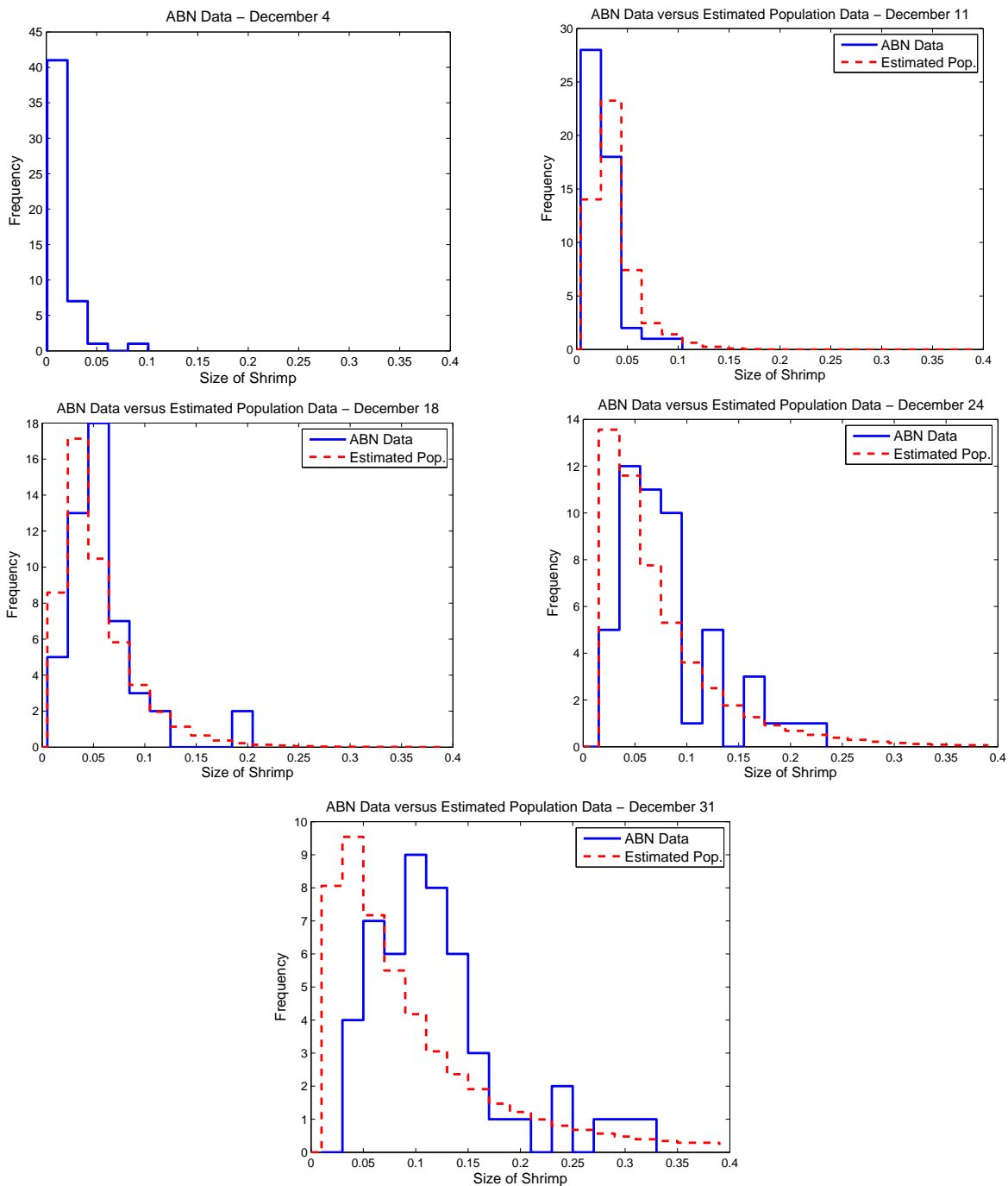


Figure 13: Longitudinal data on shrimp weight (in grams) from ABN on December 4, 11, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0018, 0.0363)$  and  $\hat{c} = 0.029$  from exponential fit of (4.7) to average size shrimp data.

## 4.2 Results Excluding December 11 Data

In an attempt to provide better fits to the shrimp population data, we excluded the data from December 11 in the parameter estimation problem for the growth rate distribution. The exponential function (4.3) generated by the deterministic growth model was used to fit the average size shrimp data for December 4, 18, 24, and 31. The estimated parameters generating the solid line shown in Figure 14 are  $\hat{a} = 0.043$ ,  $\hat{b} = 0.046$  and  $\hat{c} = 0.026$ . We note from Figure 14 that (4.3) provides a good fit to the average size shrimp data excluding the point for December 11.

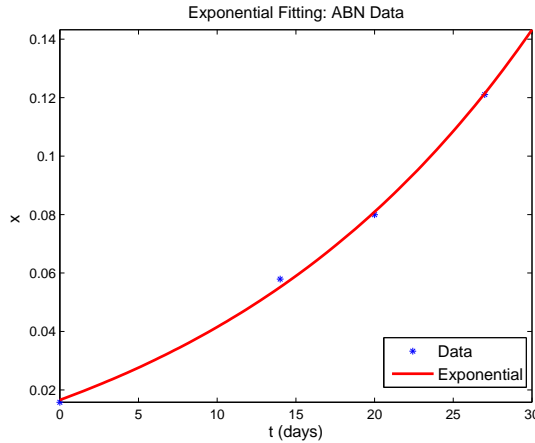


Figure 14: Exponential fit of (4.3) to ABN average size shrimp data excluding December 11 with  $\frac{d\bar{x}}{dt} = 0.046(\bar{x} + 0.026)$ .

Given the optimal value of  $c$  from the exponential fitting to averaged data, we carried out the inverse problem calculations for the mean  $\mu_b$  and standard deviation  $\sigma_b$  of the truncated normal distribution. We also interpolated the December 4 data and used it as an approximation to the initial size density  $v_0(x; g)$  in (2.1). Since we excluded the data from December 11, only the three data sets from December 18, 24, and 31 are used in estimating the growth rate distribution parameters. We obtained an optimal cost of 87.4619 and an estimated variance  $\hat{\sigma}^2$  of 3.1236 for these computations which are much smaller than the values computed in the previous calculations in Section 4.1. The estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  are in Table 6 along with the corresponding 95% confidence intervals.

$\hat{\mu}_b$	$0.0369 \pm 0.0027$	$\hat{\sigma}_b$	$0.0159 \pm 0.0030$
---------------	---------------------	------------------	---------------------

Table 6: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with ABN data excluding December 11 using  $\hat{c} = 0.026$  from exponential fit of (4.3) to average size shrimp data.

The widths of the confidence intervals for both parameters are much smaller in comparison to those computed when the entire data set is considered. We are more confident about

the estimates obtained in this case based on the relatively small confidence intervals. We were able to use the confidence intervals corresponding to the estimated mean and standard deviation of the truncated normal distribution to construct a *confidence band* or *confidence region* for the estimated probability distribution on the growth rates using techniques discussed in [9]. These results are shown in Figure 15 with the estimated probability density and confidence region on the left and the estimated probability distribution and confidence band on the right. The confidence band gives us a measure of the uncertainty associated with the infinite-dimensional parameter (probability distribution) that is of most interest in this application. Due to the width of this confidence band, we are fairly confident about the reliability of the estimated probability distribution.

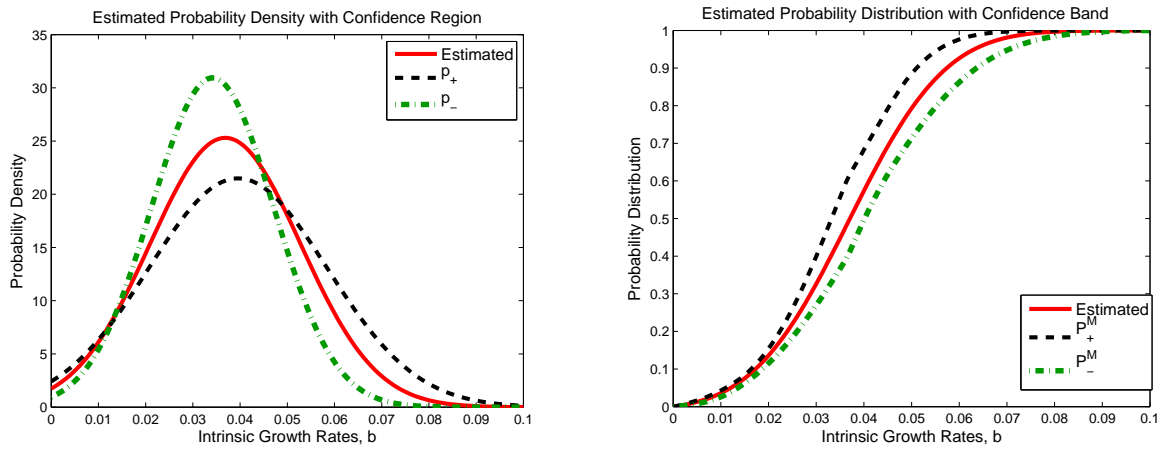


Figure 15: (left): Estimated probability density with confidence region; (right): Estimated probability distribution with confidence band for ABN data excluding December 11 using  $\hat{c} = 0.026$  from exponential fit of (4.3) to average size shrimp data.

Moreover, we note from the model predictions versus the data in Figure 16 that the model does a much better job of fitting the data on December 18, 24, and 31 when the data from December 11 is excluded. The model overall does a good job of predicting the key features of the size distribution for these three weeks.

We also used (4.7) corresponding to the probabilistic growth model (4.5) to fit the average size shrimp data with the data from December 11 excluded. The exponential fit shown in Figure 17 corresponds to the optimal parameters  $\hat{\mu}_b = 0.046$ ,  $\hat{\sigma}_b = 0.001$  and  $\hat{c} = 0.027$ . The estimated parameter values for  $\mu_b$  and  $c$  in this case are very close to the estimated parameter values of  $b$  and  $c$  when (4.3) was used to fit the data. We see that (4.7) also provides a good fit to the average size shrimp data in the absence of the December 11 data point.

The estimates for  $(\mu_b, \sigma_b)$  for the truncated normal distribution were obtained from the inverse problem calculations with the optimal value of  $c$ . The parameter estimates  $(\hat{\mu}_b, \hat{\sigma}_b)$  are relatively close to the previously estimated parameters in Table 6, and in fact, the confidence intervals in Table 7 have the same width (centered differently) as those in Table 6.

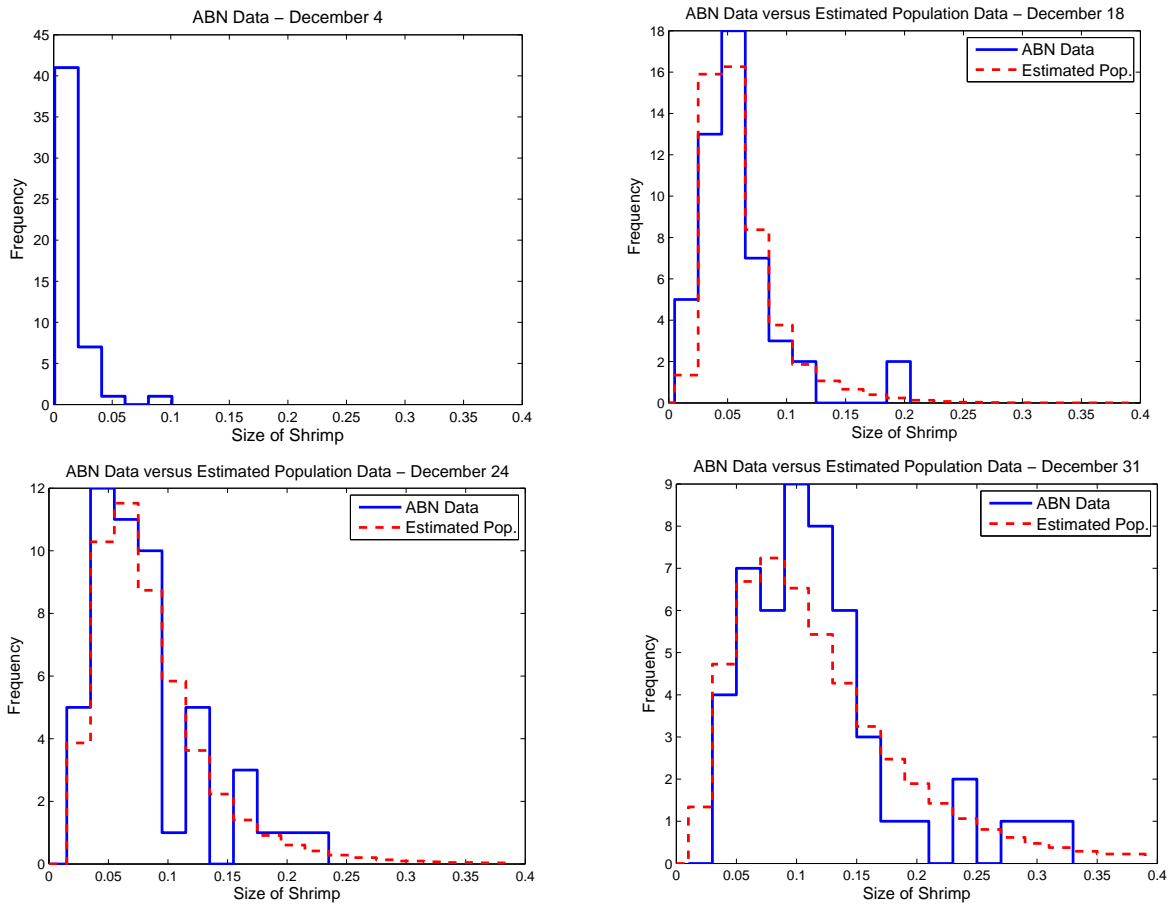


Figure 16: Longitudinal data on shrimp weight (in grams) from ABN on December 4, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0369, 0.0159)$  and  $\hat{c} = 0.026$  from exponential fit of (4.3) to average size shrimp data.

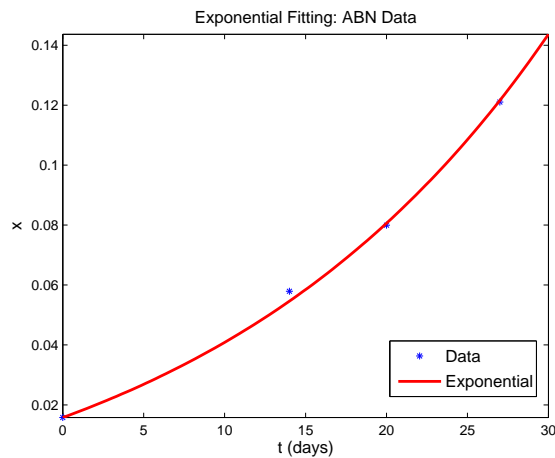


Figure 17: Exponential fit of (4.7) to ABN average size shrimp data excluding December 11 with  $\frac{d\bar{x}}{dt} = (0.046 + 0.001^2t)(\bar{x} + 0.027)$ .

$\hat{\mu}_b$	$0.0360 \pm 0.0027$	$\hat{\sigma}_b$	$0.0158 \pm 0.0030$
---------------	---------------------	------------------	---------------------

Table 7: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with ABN data excluding December 11 using  $\hat{c} = 0.027$  from exponential fit of (4.7) to average size shrimp data.

We computed an optimal cost for these results of 88.0090 and an estimated variance  $\hat{\sigma}^2 = 3.1432$ , which are also very close to the values in the previous case in this subsection. These results also support our previous assertions. The estimated parameters obtained by excluding the data from December 11 are much more reliable than those obtained when the entire data set from December is used in the parameter estimation problem. The small width observed in the confidence band associated with the estimated probability distribution in Figure 18 validate our claims as well.

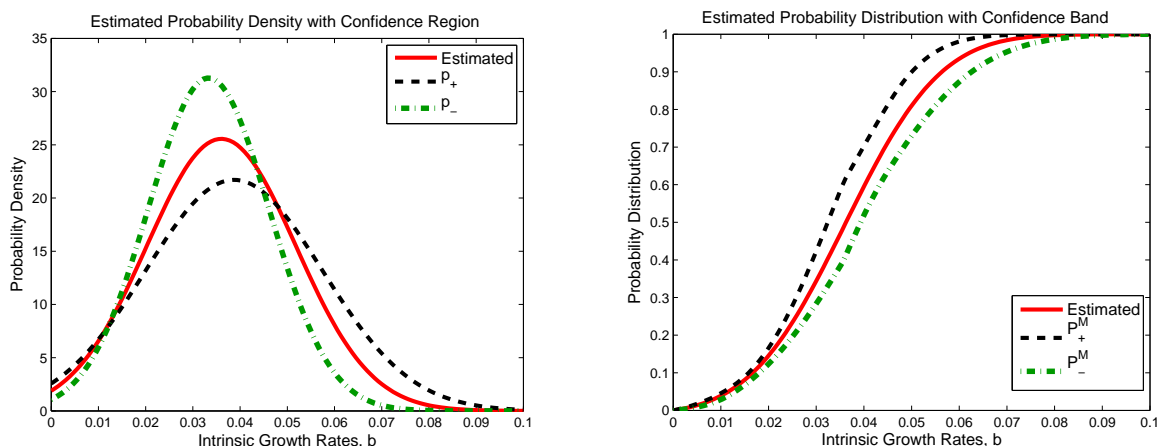


Figure 18: (left): Estimated probability density with confidence region; (right): Estimated probability distribution with confidence band for ABN data excluding December 11 using  $\hat{c} = 0.027$  from exponential fit of (4.7) to average size shrimp data.

In Figure 19 we see similar results to those shown in Figure 16 with there being a very small difference in the model predictions on December 18. As previously remarked, the model does a very good job of fitting the data for December 18, 24, and 31 when the data from December 11 is excluded from the inverse problem calculations. The model predictions exhibit the lognormal behavior seen in the observations and provide reasonably accurate quantitative fits to the data.

While we were able to obtain reliable parameter estimates and good model predictions by excluding the data from December 11, we also chose to investigate the effects of the initial data in this problem by excluding the data from December 4 and using the data from December 11 as an approximation to the initial size density  $v_0(x; g)$ . The results in this section suggest that perhaps there was a change in the environmental conditions, such as feeding or temperature, which affected the growth of the shrimp. However, the poor model fits in Section 4.1 where the entire data set from December is considered could also be the result of using unreliable initial conditions. Therefore, we reconsidered the parameter

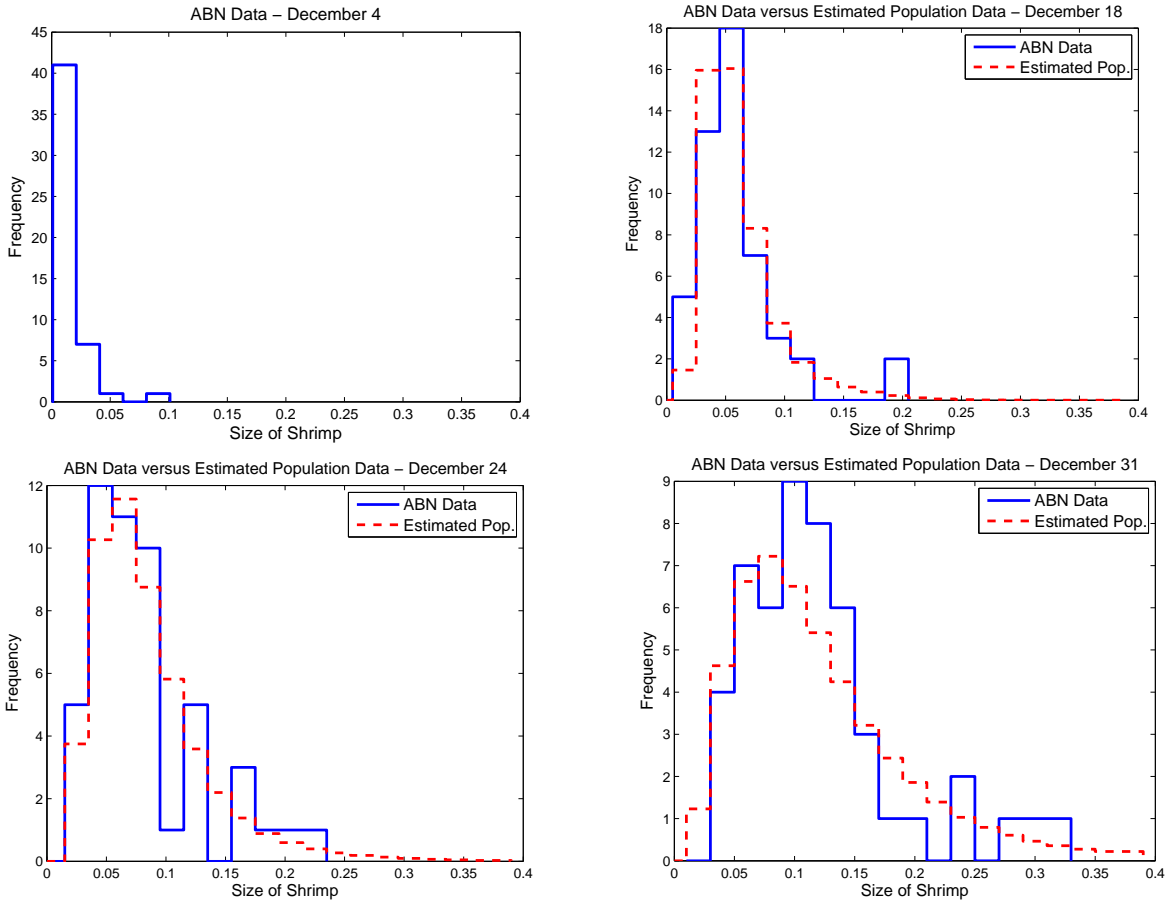


Figure 19: Longitudinal data on shrimp weight (in grams) from ABN on December 4, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0360, 0.0158)$  and  $\hat{c} = 0.027$  from exponential fit of (4.7) to average size shrimp data.

estimation problem for the growth rate distribution with the data collected on December 11, 18, 24, and 31, and the results follow in the next section.

### 4.3 Results Excluding December 4 Data

We now present results for the inverse problem for the growth rate distribution of the size-structured shrimp population considering the data from December 11 - 31 in order to determine if using the interpolated December 11 data as an approximation for  $v_0(x; g)$  will result in better model predictions. Under the assumption of the deterministic growth model (4.4), Equation (4.3) was used to fit the average size shrimp data excluding the data point for December 4. The exponential function in (4.3) with the optimal parameter values of  $\hat{a} = 0.058$ ,  $\hat{b} = 0.047$ ,  $\hat{c} = 0.028$  fits the average size shrimp data well as shown in Figure 20.

The inverse problem for  $(\mu_b, \sigma_b)$  was performed with the optimal value of  $c$  computed from

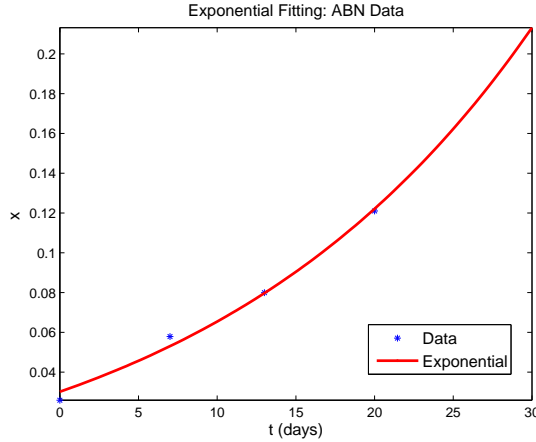


Figure 20: Exponential fit of (4.3) to ABN average size shrimp data excluding December 4 with  $\frac{d\bar{x}}{dt} = 0.047(\bar{x} + 0.028)$ .

the exponential fitting to the average size shrimp data. As already noted, we interpolated the data from December 11 and used it as an approximation to the initial size density. We were left with the final three data sets from December 18, 24, and 31 for the parameter estimation problem. The optimal parameter values for the mean and standard deviation of the truncated normal distribution along with the computed confidence intervals are in Table 8. We note that the optimal cost  $J(q^*) = 121.4561$  and the estimated variance of the statistical model  $\hat{\sigma}^2 = 4.3377$ . Although these values are larger than the values obtained when the data from December 4 was used as an approximation to the initial condition, the results are still significantly better than those obtained with the entire data set.

$\hat{\mu}_b$	$0.0441 \pm 0.0077$	$\hat{\sigma}_b$	$0.0294 \pm 0.0123$
---------------	---------------------	------------------	---------------------

Table 8: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with ABN data excluding December 4 using  $\hat{c} = 0.028$  from exponential fit of (4.3) to average size shrimp data.

We also observe slightly larger confidence intervals corresponding to the estimated parameters  $\hat{\mu}_b$  and  $\hat{\sigma}_b$ . These larger confidence intervals result in the larger confidence band for the estimated probability distribution in Figure 21. Thus, there is more uncertainty in the estimates obtained in this case in comparison to the estimates obtained when the data from December 11 was excluded in Section 4.2. We are more confident about the reliability of the estimates obtained for  $(\mu_b, \sigma_b)$  in Section 4.2 based on the smaller width of the confidence intervals and bands.

Although the optimal cost was slightly larger in this case, the model predictions shown in Figure 22 are similar to those in Section 4.2. The model still does a good job of predicting the key features of the shrimp population data when the data from December 11 is used

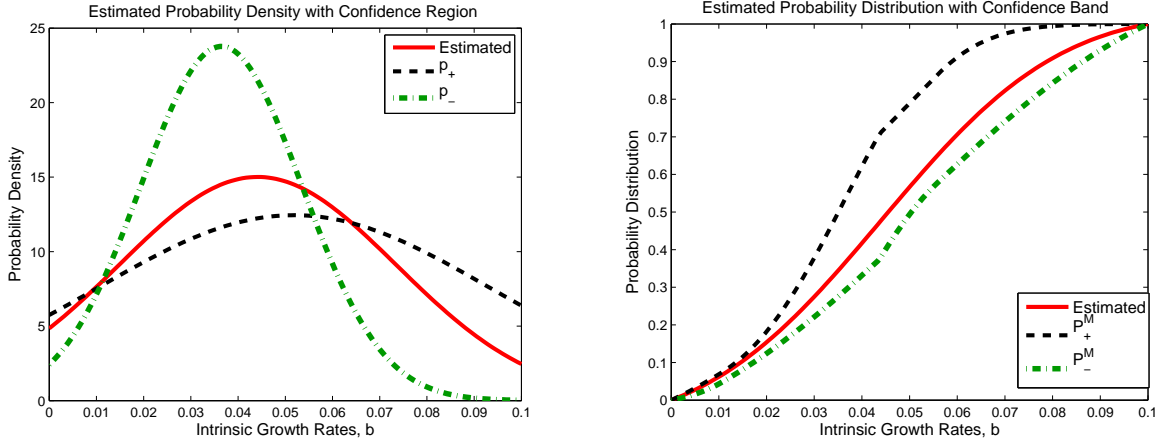


Figure 21: (left): Estimated probability density with confidence region; (right): Estimated probability distribution with confidence band for ABN data excluding December 4 using  $\hat{c} = 0.028$  from exponential fit of (4.3) to average size shrimp data.

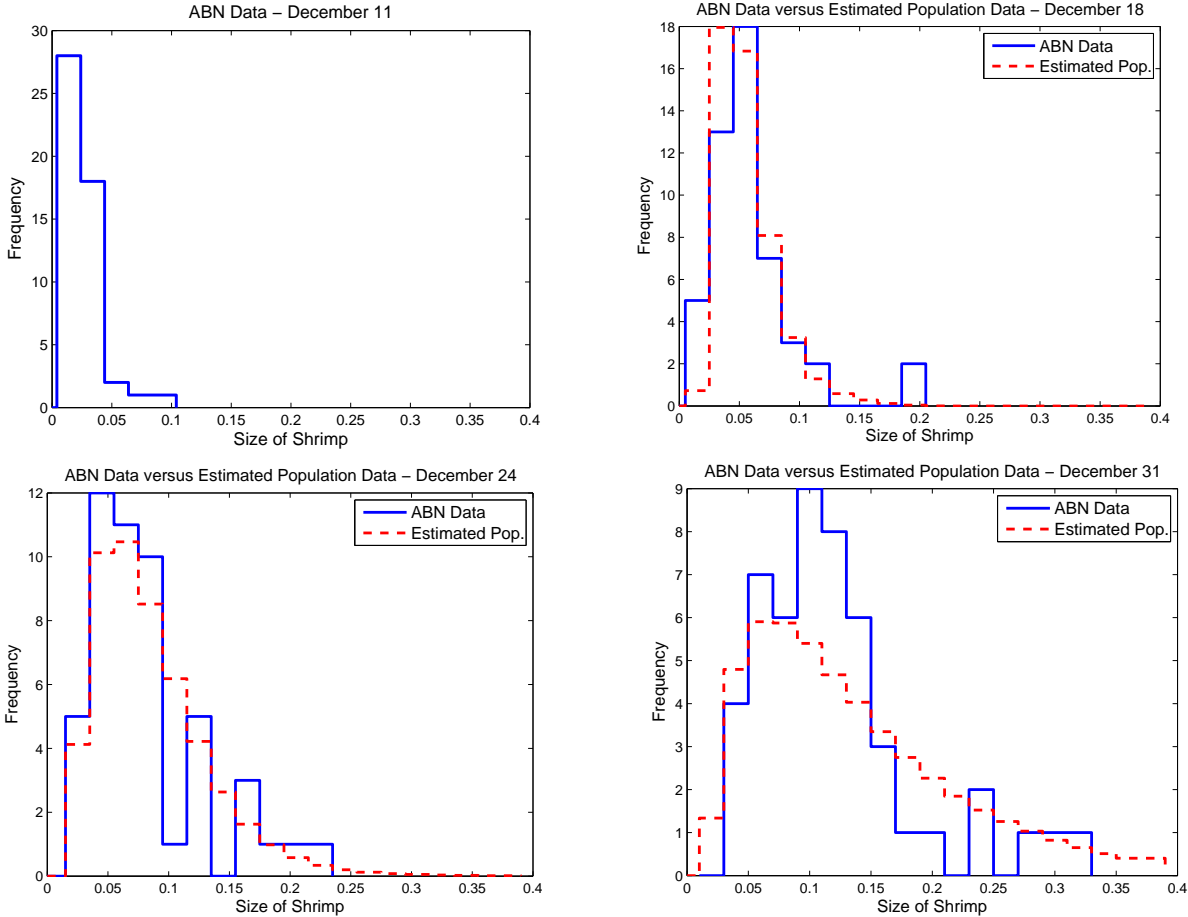


Figure 22: Longitudinal data on shrimp weight (in grams) from ABN on December 11, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0441, 0.0294)$  and  $\hat{c} = 0.028$  from exponential fit of (4.3) to average size shrimp data.



versus the data from December 4. It appears that using the December 11 data instead of the December 4 data does not result in discernable improvements in the model fits to data.

In our final set of computations, we fit the average size shrimp data with the data point from December 4 excluded with (4.7) taking into account the probabilistic nature of  $b$ . There is some noticeable difference in the estimated parameter values  $\hat{\mu}_b = 0.038$ ,  $\hat{\sigma}_b = 0.003$ ,  $\hat{c} = 0.059$  determined to give the best fit shown in Figure 23. However, we note that the exponential fit in Figure 23 is comparable to that shown in Figure 20.

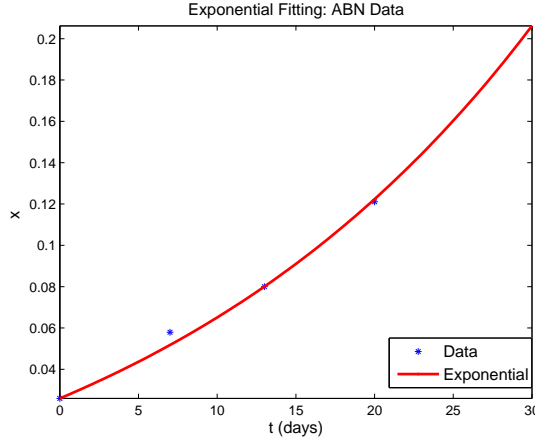


Figure 23: Exponential fit of (4.7) to ABN average size shrimp data excluding December 4 with  $\frac{d\bar{x}}{dt} = (0.038 + 0.003^2t)(\bar{x} + 0.059)$ .

The results of the inverse problem calculations for the mean  $\mu_b$  and standard deviation  $\sigma_b$  of the truncated normal distribution are in Table 9 along with the confidence intervals. We computed an optimal cost of 106.6413 and an estimated variance  $\hat{\sigma}^2$  of 3.8086 which are slightly better than the previous results in this section. We point out that the confidence intervals for the estimated mean in Table 9 do not intersect with those in Table 8, but the confidence intervals for  $\hat{\sigma}_b$  do intersect. However, we still infer from the results shown here that the parameter estimates obtained in absence of the data from December 4 are not as reliable as those obtained in the previous section where the data from December 11 was omitted. The confidence band associated with the estimated probability distribution in Figure 24 also suggests that we are not as confident about the reliability of this estimated probability distribution in comparison to the estimates presented in Section 4.2.

$\hat{\mu}_b$	$0.0289 \pm 0.0051$	$\hat{\sigma}_b$	$0.0215 \pm 0.0065$
---------------	---------------------	------------------	---------------------

Table 9: Estimated parameters  $(\hat{\mu}_b, \hat{\sigma}_b)$  and confidence intervals for inverse problem with ABN data excluding December 4 using  $\hat{c} = 0.059$  from exponential fit of (4.7) to average size shrimp data.

The model predictions in Figure 25 are very similar to the previous results. We observe

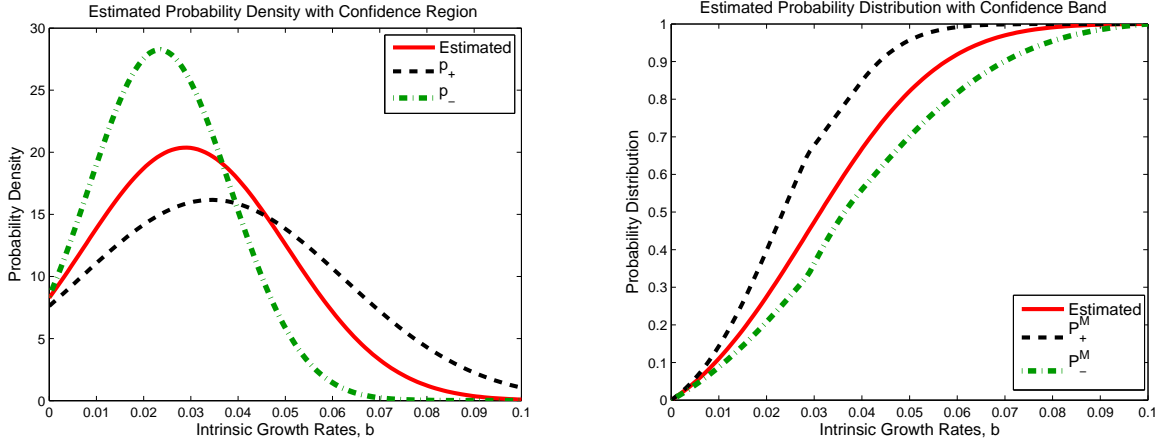


Figure 24: (left): Estimated probability density with confidence region; (right): Estimated probability distribution with confidence band for ABN data excluding December 4 using  $\hat{c} = 0.059$  from exponential fit of (4.7) to average size shrimp data.

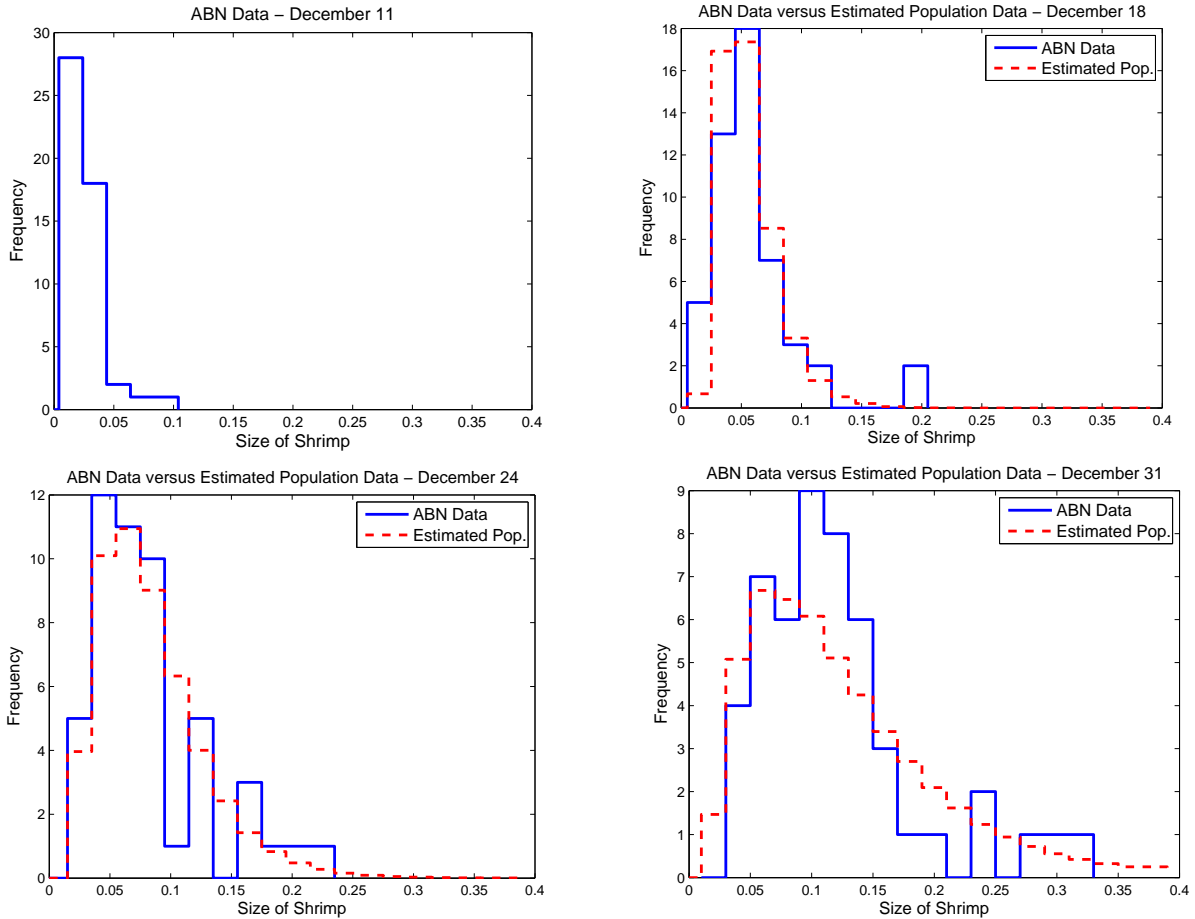


Figure 25: Longitudinal data on shrimp weight (in grams) from ABN on December 11, 18, 24, and 31 versus model predicted population density  $p(x, t; q^*)$  with  $q^* = (0.0289, 0.0215)$  and  $\hat{c} = 0.059$  from exponential fit of (4.7) to average size shrimp data.

nice model fits to the data which validates the model's ability to accurately describe the growth dynamics of the shrimp. The results presented in this section suggest that there were perhaps some changes in the environmental conditions between December 4 and December 11 which resulted in very poor model predictions when the entire data set was used in the parameter estimation problem. Although we were able to use the data from December 11 as an approximation to the initial size density to compute good fits to the data, the best numerical and statistical results were obtained when the data from December 11 was excluded from the inverse problem calculations. These results are illustrative of the types of issues that can be addressed with mathematical and statistical inverse problem ideas when using models to describe specific applications.

## 5 Concluding Remarks

In this paper we have presented a computational methodology based on statistical and mathematical estimation and inverse problem techniques. This methodology can be used to design experiments that can provide data adequate to reliably estimate parameters in partial differential equation models for growth rates in populations. We presented this in the context of a Growth Rate Distribution (GRD) model for size-structured shrimp populations whose size distribution data exhibited a great deal of variability. We then detailed the computational methodology that proved to be useful in designing experiments that were carried out at ABN to validate the GRD model. We were able to determine the *sampling size* and *frequency* sufficient for reliable estimates of the variability in the growth rates of size-structured shrimp populations. Using our inverse problem methodology, we were able to successfully estimate the growth rate distribution from experimental data (from experiments based on our design calculations) collected during the early growth process of shrimp populations at ABN. The results presented in this paper demonstrate how the GRD model can be used to gain insight into the growth processes of shrimp. Moreover, although the mathematical and statistical methodologies were presented in the context of size-structured shrimp populations, this approach and these ideas can easily be extended to other problems in which one wishes to design and use experimental data in model investigation and validation.

## Acknowledgments

The authors are most grateful to Dr. Tzachi M. Samocha of the Shrimp Mariculture Research Facility, Texas Agricultural Experiment Station, 4301 Waldron Road, Corpus Christi, TX 78418, for providing the juvenile shrimp size distribution data presented in Figures 1 and 2 in this manuscript. This research was supported in part by the National Institute of Allergy and Infectious Disease under grant 9R01AI071915-05, in part by the U.S. Air Force Office of Scientific Research under grants AFOSR-FA9550-04-1-0220 and AFOSR-FA9550-08-1-0147, in part by the Defense Advanced Research Projects Agency and Advanced BioNutrition

Corp under grant ABNDTRA-0507-03 and in part by the US Department of Energy Computational Science Graduate Fellowship to J.L. Davis under grant DE-FG02-97ER25308.

## References

- [1] L.J.S. Allen, *An Introduction to Stochastic Processes with Applications to Biology*, Pearson Education, Inc., Upper Saddle River, New Jersey, 2003.
- [2] H.T. Banks, J.E. Banks, L.K. Dick, and J.D. Stark, Estimation of dynamic rate parameters in insect populations undergoing sublethal exposure to pesticides, CRSC-TR05-22, NCSU, May, 2005; *Bulletin of Mathematical Biology*, **69** (2007), 2139–2180.
- [3] H.T. Banks, J.E. Banks, L.K. Dick, and J.D. Stark, Time-varying vital rates in ecotoxicology: selective pesticides and aphid population dynamics, *Ecological Modelling*, **210** (2008), 155–160.
- [4] H.T. Banks and K.L. Bihari, Modeling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111.
- [5] H.T. Banks, V.A. Bokil, S. Hu, F.C.T. Allnut, R. Bullis, A.K. Dhar and C.L. Browdy, Shrimp biomass and viral infection for production of biological countermeasures, CRSC-TR05-45, NCSU, December, 2005; *Mathematical Biosciences and Engineering*, **3** (2006), 635–660.
- [6] 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 Mathematics **FR28**, SIAM, Philadelphia, 2003, pp. 129–154.
- [7] H.T. Banks, L.W. Botsford, F. Kappel, and C. Wang, Modeling and estimation in size structured population models, LCDS-CCS Report 87-13, Brown University; *Proceedings 2nd Course on Mathematical Ecology*, (Trieste, December 8-12, 1986) World Press (1988), Singapore, 521–541.
- [8] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, CRSC-TR05-38, NCSU, October, 2005; *Applied Numerical Mathematics*, **57** (2007), 753–777.
- [9] H.T. Banks and J.L. Davis, Quantifying uncertainty in the estimation of probability distributions with confidence bands, CRSC-TR07-21, NCSU, December, 2007; *Mathematical Biosciences and Engineering*, **5** (2008), 647–667.
- [10] H.T. Banks, J.L. Davis, and S. Hu, Comparison of Fokker-Planck model and Growth Rate Distribution model in modeling growth uncertainty (in progress).

- [11] H.T. Banks, J.L. Davis, S.L. Ernstberger, S. Hu, E. Artimovich, A.K. Dhar, C.L. Browdy, A comparison of probabilistic and stochastic formulations in modeling growth uncertainty and variability, CRSC-TR08-03, NCSU, February, 2008; *Journal of Biological Dynamics* (to appear).
- [12] H.T. Banks and B.G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, University of Southern California, January, 1990; *Quarterly of Applied Mathematics*, **49** (1991), 215–235.
- [13] H.T. Banks, B.G. Fitzpatrick, L.K. Potter, and Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data, CRSC-TR98-6, NCSU, January, 1998; In *Stochastic Analysis, Control, Optimization and Applications*, (Edited by W. McEneaney, G. Yin and Q. Zhang), Birkhäuser, Boston, 1989.
- [14] H.T. Banks and G.A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, CRSC-TR04-03, NCSU, January, 2004; *SIAM J. Multi-scale Modeling and Simulation*, **2** (2005), 395–412.
- [15] H.T. Banks, H.T. Tran and D.E. Woodward, Estimation of variable coefficients in Fokker Planck equations using moving finite elements, CAMS Tech. Rep. 90-9, August, 1990, University of Southern California; *SIAM Journal of Numerical Analysis*, **30** (1993), 1574–1602.
- [16] G.I. Bell and E.C. Anderson, Cell growth and division I. A mathematical model with application to cell volume distributions in mammalian suspension cultures, *Biophys. J.*, **7** (1967), 330–351.
- [17] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
- [18] G. Casella and R.L. Berger, *Statistical Inference*, Duxbury, California, 2002.
- [19] F.L. Castille, T.M. Samocha, A.L. Lawrence, H. He, P. Frelier, and F. Jaenike, Variability in growth and survival of early postlarval shrimp (*Penaeus vannamei* Boone 1931), *Aquaculture*, **113** (1993), 65–81.
- [20] M. Davidian and A.R. Gallant, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1995.
- [21] A.R. Gallant, *Nonlinear Statistical Models*, J. Wiley & Sons, New York, 1987.
- [22] T.C. Gard, *Introduction to Stochastic Differential Equations*, Marcel Dekker, New York, 1988.
- [23] A. Gross, S. Abutbul, and D. Zilberg, Acute and chronic effects of nitrite on white shrimp, *Litopenaeus vannamei*, cultured in low-salinity brackish water, *Journal of the World Aquaculture Society*, **35** (2004), 315–321.
- [24] R.I. Jennrich, Asymptotic properties of non-linear least squares estimators, *The Annals of Mathematical Statistics*, **40** (1969), 633–643.

- [25] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, 2001.
- [26] S. Laramore, C.R. Laramore, and J. Scarpa, Effect of low salinity on growth and survival of postlarvae and juvenile *litopenaeus vannamei*, *Journal of the World Aquaculture Society*, **32** (2001), 385–392.
- [27] T. Luzyanina, D. Roose, T. Schinkel, M. Sester, S. Ehl, A. Meyerhans and G. Bocharov, Numerical modeling of label-structured cell population growth using CFSE distribution data, *Theor. Biol. and Med. Model.*, **4** (2007), 1–26.
- [28] J.A.J. Metz and O. Diekmann, *The Dynamics of Physiologically Structured Populations*, Lec. Notes in Biomath, Vol. **68**, Springer-Verlag, Berlin, 1986.
- [29] A. Okubo, *Diffusion and Ecological Problems: Mathematical Models*, Biomathematics, Vol. **10**, Springer-Verlag, Berlin, 1980.
- [30] J. Ponce-Palafox, C.A. Martinez-Palacios, and L.G. Ross, The effects of salinity and temperature on the growth and survival rates of juvenile white shrimp, *penaeus vannamei boone 1931*, *Aquaculture*, **157** (1997), 107–115.
- [31] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley & Sons, New York, 1989.
- [32] J.W. Sinko and W. Streifer, A new model for age-size structure for a population, *Ecology*, **48** (1967), 910–918.