

Analysis of Panel Data via Mechanistic Models in PanelPOMP Framework

A Thesis Submitted in Partial Fulfillment of Bachelor of Science in
Honor Data Science at the University of Michigan

BO YANG

Thesis Advisor

Professor Edward Ionides

Jesse Wheeler

Professor Aaron King

Department of Statistics, University of Michigan

Apr 2023

Abstract

In ecological research, panel data is a typical format of data storage. Panel data is a collection of time series, where each time series corresponds to an individual or unit in a system. As a result of its high dimension, panel data analysis presents significant difficulties. Mechanistic modeling is an attractive approach to ecological modeling, but is difficult to conduct in high-dimensions. Mechanistic models enable quantitative analysis, which provides a deeper understanding of the latent process of the dynamic system, and provide inference and prediction based on panel data. The objective of this study is to showcase the practicality and advancement of mechanistic model development for ecological panel datasets using the panelPOMP structure and panel iterated filtering (PIF). PIF is used as a probability maximization and parameter estimation technique. We propose three different ecological models to panel *Daphnia* data, and fit each model via maximum likelihood. The best model is selected using Akaike information criterion (AIC). Our results suggest that alga, a major food source for *Daphnia*, plays an important role in *Daphnia* population dynamics. We illustrate the effectiveness of a panelPOMP framework in modeling the complex nature of ecological data analysis, emphasizing its capacity to provide meaningful explanations and in-depth understanding of latent processes not captured by existing data. A discussion on the limitations and interpretations of mechanistic models in dealing with ecological data is included.

1 Introduction

Ecological experiments may contain many elements or processes that are unobservable, models that can describe both observable and unobservable processes present in a dynamic system of great interest. Traditional time series analysis offers limited explanatory power due to the nonstationary, nonlinear, and stochastic nature of ecological information. As a mathematical description of the elements forming a system, mechanistic models can describe the mutual interactions between elements in the system and their environment (Stalidzans et al., 2020). By accounting for known biological phenomena, mechanistic models elucidate the actions of hidden yet important mechanisms (Duarte et al., 2003). Mechanistic modeling involves employing mathematical models to describe the development and interaction of dynamic systems, allowing researchers to evaluate the explanatory potential of various conceptual models and measure the magnitude of pertinent parameters. Mechanistic modeling has been widely employed in social sciences (Holme and Liljeros, 2015). Improved computational capabilities now make it more effective. Despite the creation of theoretical models for the data, connecting theories with data continues to be a challenging task. As a result, ecological mechanical modeling cannot solve many of the problems on a wide scale, but rather, only a limited range of the problems (Duarte et al., 2003).

Panel data, sometimes referred to as longitudinal data, is composed of multiple time series. Each of these time series, which may be multivariate, contains a series of observations collected from a separate entity (Bretó et al., 2020). Panel data is common in ecology due to the nature of experiments in the field; data are often collected over time for multiple entities that represent various degrees of treatment and control. Due to the non-linear characteristics of ecological data, the analysis of panel data entails a non-linear and high-dimensional structure. Such features diminish the explanatory power of traditional time series analysis and certain Monte Carlo inference techniques. The goal of this study is to investigate the practicality of employing mechanistic models on ecological panel data using the panelPOMP structure.

The PanelPOMP approach is a special case of the partially observable Markov Process (POMP) model to accommodate panel data by constructing a POMP model for each individual entity. To fit both shared and unit specific parameters, we use the panel iterated filtering (PIF) algorithm (Bretó et al., 2020), which calibrates parameters via maximum likelihood. Based on iterated filtering (King et al., 2016), an algorithm applied to POMP models of a single unit, panel iterated filtering not only filters within each series but also cycles across the panel to achieve the best likelihood. This enables the application of mechanistic models to panel data while preserving dimensionality.

In this research, we will demonstrate the usefulness of the panelPOMP framework by analyzing the *Daphnia* panel data collected in the experiment (Searle et al., 2016). The panel data describes the change in population densities of different species of *Daphnia* under different conditions over time. One goal of their study was to interpret how the presence of invasive *Daphnia* species (*lumholtzi*) affects the response of native *Daphnia* species (*dentifera*) to parasites (Searle et al., 2016). In their lab experiment, they generated standard ecological panel data featuring various independent trials

under the same experimental conditions, with each resulting in a brief, nonlinear, and nonstationary time series. The nonlinear nature of the data makes it impractical to use many conventional time series methods. Utilizing the panelPOMP approach allows for the analysis of these data and a mechanistic interpretations of the results, enabling the extraction of the information found in each individual time series.

For a more in-depth examination of the latent processes in the panel data, this research incorporates two classic ecological models—the Logistic Growth model and the Lotka-Volterra model—within the panelPOMP framework. After proposing both models to the *Daphnia* panel data, parameter estimates are calculated via the PIF algorithm. The resulting models are subsequently compared in order to discover a more accurate and appropriate model for the given data. The superior performance of the Lotka-Volterra model implies that the population density of alga *Ankistrodesmus falcatus* (Searle et al., 2016), which is *Daphnia*'s prey, impacts the dynamics of both species of *Daphnia*. It is worth noting that the population density of alga is neither observable nor recorded in the experiment, making the panelPOMP framework a desirable approach since it accounts for this important latent processes. By analyzing *Daphnia* panel data, we hope to show the limitations and capabilities of mechanistic models.

2 Data description

2.1 *Daphnia*-alga Data

Searle and colleagues conducted two experiments to investigate the interaction between native and invasive *Daphnia* species and a parasitic fungus called *Metschnikowia bicuspidate* (Searle et al., 2016). The first experiment was an individual-level study that compared the host competence of the two *Daphnia* species. The second experiment was a mesocosm experiment that examined the community-level species interaction and the role of population density in disease patterns. The mesocosm experiment involved six different treatments, including both single-species and mixed-species treatments, with and without parasites. The experiment was conducted indoors, with 45 *Daphnia* added to each 15L high-hardness COMBO media, which is a defined freshwater culture medium for algae and zooplankton, (Baer and Goulden, 1998), at the start of the four single-species treatments and 10 invasive and 35 host species added in the treatments of mixed-species. The same treatment will be repeated for ten times, and samples were taken every five days for 52 days after the experiment began. During the experiment, 1.3×10^6 cells L^{-1} of algae *falcatus* will be added twice a week. Measurements were obtained by sampling one liter of COMBO solution each week for all ten units. The genus, disease status, generation stage, and sex of each individual were documented, and the temperature in the lab was kept constant at 23.27°C (SE \mp 0.2) with a 16L:8D photoperiod and all the other conditions remained the same for all units during the experiment time.

One question of interest in this experiment is the interplay between the *Daphnia* species (*Daphnia dentifera* and *Daphnia lumholtzi*) and their surroundings in the absence of parasites. In order to address this question, we analyze a panel data con-

sisting of ten independent time series representing the population density of *dentifera* adults. These observations were gathered simultaneously at ten time points, as shown in Figure 1 and 2, and offer the opportunity to evaluate the relationship between alga and *Daphnia*, and compare mechanistic models that describe that relationship.

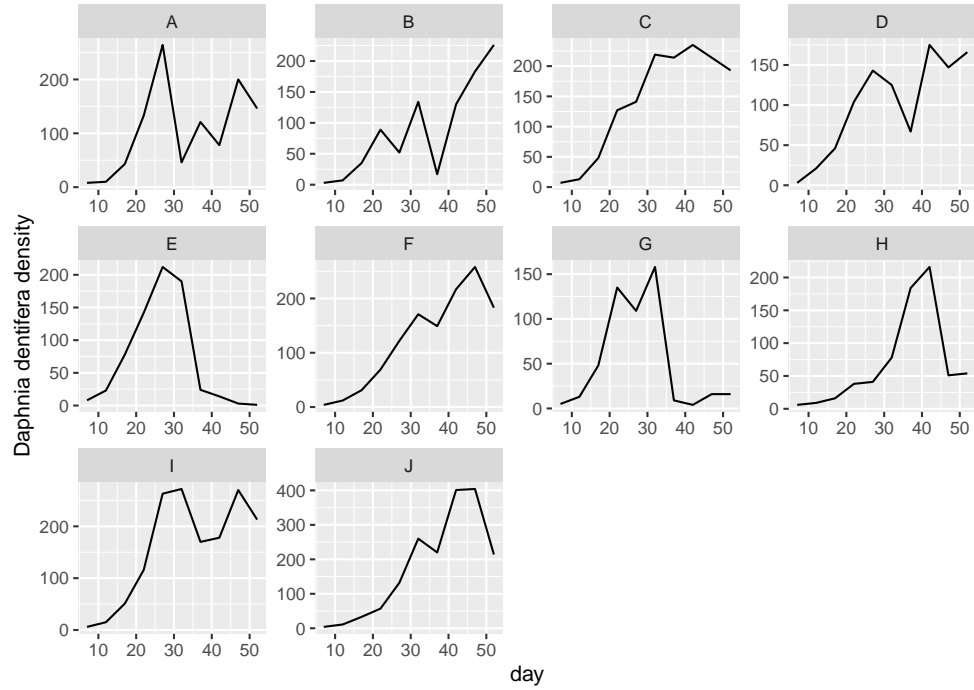


Figure 1: Data visualization for *Daphnia dentifera*: Ten replications of *Daphnia* population density panel

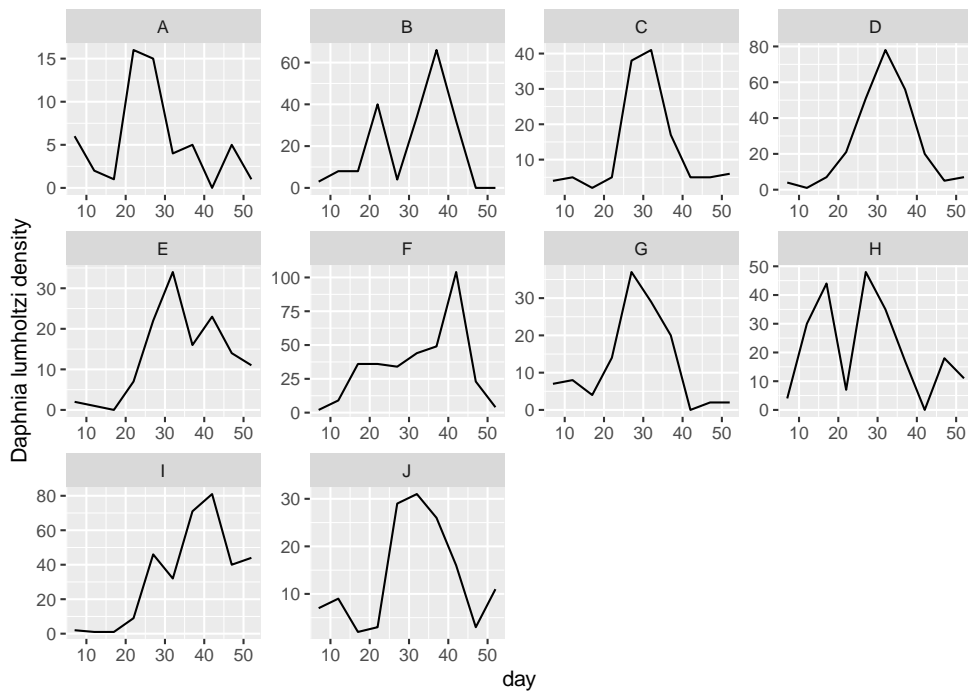


Figure 2: Data visualization for *Daphnia lumholzi*: Ten replications of *Daphnia* population density panel

3 Methodology

3.1 Partially Observed Markov Processes

Frequently, we lack a comprehensive understanding of the underlying mechanisms that drive the evolution of a natural system. For example, the precise number of individuals who are exposed to a disease or the density of alga at any given time is typically unknown. To address these issues of missing information, we can utilize partially observed Markov (POMP) models. These models describe how the latent states of a system change over time and how measurements on the system are obtained.

With the random variables $Y_{0:N}$ which are the observable states process, the measurement model is given by $f_{Y_n|X_n}(y_n|x_n; \theta)$; the observed data $y_{0:N}$ are assumed to be random draws from this density. We denote the latent states as the random variable $X_{0:N}$, where X_0 represents the initial state of the latent process. The evolution of the latent state is described as samples from a one-step transition density $f_{X_n|X_{n-1}}(x_n|x_{n-1})$. We assume that at each time point $n \in 1 : N$, Y_n depends only upon the latent process X_n , and is conditionally independent of Y_{n-1} and X_{n-1} .

3.2 Likelihood of models and PIF

In statistical inference, the model likelihood frequently serves as the basis for estimating parameters and selecting models. Considering a model parameterized by vector θ within the m -dimensional parameter space Θ_m , the likelihood function represents

the combined probability density of the dataset, $y_{1:N}$, at θ :

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}; \theta) \quad (1)$$

Because elements of θ are typically unknown, estimates of θ are obtained by maximizing the likelihood function. Any parameter vector that maximizes the likelihood function is called a maximum likelihood estimate, and is denoted $\hat{\theta}$.

An ecological model's effectiveness in tracking disease spread hinges on its capability to predict cases or incidence rates. This capability is inherently tied to the model's ability to accurately describe the temporal changes in latent states. As a result, we encounter two interrelated issues: finding plausible values of X_n at time n based on $y_{1:n}$, and pinpointing parameter values, $\hat{\theta}$, that maximize the model likelihood. These processes are commonly referred to as the filtering process and the inference process, respectively.

The nonlinear nature of arbitrary POMP models often makes analytical evaluation of the likelihood function and transition densities impossible. Simulation based methods such as the particle filter (Arulampalam et al., 2002) help address this issue by enabling the calculation of likelihood values while only requiring the ability to simulate from the model and evaluate the measurement model.

To implement the particle filter, we start by initializing M particles at time t_0 , where $\{X_0^m; m \in 1 : M\}$. Then, for every time step $n \in 1 : N$, we advance the each particle by simulating from the transition density, resulting in a set of particles that represents the prediction distribution at time n , denoted as $X_n^{m,P} \sim f_{X_n|X_{n-1}}(\cdot|X_{n-1}^m; \theta)$, for $m \in 1 : M$. Next, we assign weights to the particles based on the measurement density to obtain an ensemble that represents the filtering distribution at time n , where the new weight is calculated via $w_{n,m} = f_{Y_n|X_n}(y_n|X_n^{m,P}; \theta)$. Finally, we redraw particles based on their weights to generate a new set of particles that represents the filtering distribution, denoted as $X_n^{m,F}$. If the number of particles is large, we can approximate the likelihood using equations (1) and (4):

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}; \theta) \quad (2)$$

$$= \prod_{n=1}^N f_{Y_n|Y_{1:n-1}}(y_n|y_{n-1}; \theta) \quad (3)$$

$$= \prod_{n=1}^N f_{Y_n|X_n}(y_n|x_n; \theta) \quad (4)$$

$$\approx \prod_{n=1}^N \frac{1}{M} \sum_{m=1}^M f_{Y_n|X_n}(y_n|x_n^m) \quad (5)$$

This algorithm offers a straightforward approach to estimate the likelihood and the latent state at a specific time based on the available observations (Ionides et al., 2015).

3.3 PanelPomp Framework

In a panel data set containing K independent time series, K entities are represented by the symbols u_1, u_2, \dots, u_K , with N_k periodic observations collected for each unit. A stochastic observable process, denoted as $y_{k,1}, \dots, y_{k,N}$, represents measurements taken

on the system. PanelPOMP is a special case of the POMP model, where the latent and observable processes can be factored into independent entities (Bretó et al., 2020), and entails developing a POMP model for each entity. We use the PanelPOMP structure to describe three ecological models for a single-species experimental container without parasites.

4 Models

In this section, we describe three models that are used to fit the *Daphnia* panel data. The first two provide only simple representations of the system and are used to benchmark the results of the third model.

4.1 Logistic Model

The simplest of the models is the Logistic Model, where the *Daphnia* Population Density at time t is modeled using the state variable $S(t)$, and the population changes according to the following equations:

$$dS(t) = (r + \zeta) S(t) \left(1 - \frac{S(t)}{K}\right) \cdot dt - \delta S(t) \cdot dt \quad (6)$$

$$\zeta \sim N(0, \sigma_G) \quad (7)$$

This model contains one state variable *Daphnia* population density (S) and three parameters: r, K and σ_G . r represents the growth rate of *Daphnia*; K is the carrying capacity of this ecosystem in terms of density of *Daphnia*; and δ denotes the sampling rate in the experiment. In order to minimize the dimension of parameters, we fixed the sample rate δ as a constant equal 0.013 (Searle et al., 2016). ζ is a noise term added to population growth to allow for random variation in the dynamic. In this case, we assure ζ follows a Normal distribution with 0 mean and a standard deviation of σ_G .

The logistic model is a well-known ecological model that is commonly used to describe the population growth of a species under limited conditions, such as space or food availability (Lomnicki, 1980). Given that the population growth curve appears to follow a similar pattern to the logistic model, we attempt to fit the data to this model. As the population density (S) approaches the carrying capacity (K), the rate of population growth decreases and becomes zero when S equals K . If S exceeds K , the population will decline. Moreover, as S increases, the noise in the data also increases, which will be accounted for by the model.

This model implicitly assumes that the population has sufficient food, and that the limiting factor for growth is the space available for *Daphnia*. Therefore if the model fits the data accurately, it would imply that the population has sufficient food and that the limiting factor is the space available for *Daphnia*. However, if the model does not fit the data well, then we should consider the inclusion of alga into the model, as algae are a major food source for *Daphnia*, and a potentially significant component of the population dynamics.

4.2 Lotka-Volterra model with enough food supply

The next model includes a new latent state F that models the native supply of algae in each unit (Yang, 2018):

$$dF(t) = \alpha F(t) \left(1 - \frac{F(t)}{k_f}\right) \cdot dt + F(t) \left(1 - \frac{F(t)}{k_f}\right) d\zeta_F - \beta F(t) S(t) \cdot dt \quad (8)$$

$$dS(t) = \theta S(t) F(t) \cdot dt + S(t) F(t) d\zeta_S - \gamma S(t) \cdot dt - \delta S(t) \cdot dt \quad (9)$$

$$d\zeta_F \sim N(0, \sigma_F^2 \cdot dt) \quad (10)$$

$$d\zeta_S \sim N(0, \sigma_S^2 \cdot dt) \quad (11)$$

This model includes eight parameters for each unit: $\gamma, \alpha, \sigma_F, \sigma_S, \theta, \beta, \delta$ and k_f . α represents the rate of growth of algae; k_f shows the carrying capacity of algae in each bucket; β is the consumption rate of algae by certain species of *Daphnia*; θ shows the rate of growth of *Daphnia* with certain type of food; and γ is the rate of death of certain species of *Daphnia*. Similar to the logistic model, we treat δ as a constant equal to 0.013, according to the experimental settings (Searle et al., 2016). $d\zeta_F$ and $d\zeta_S$ are Brownian noise terms of algae (F) growth and *Daphnia* (S) growth in order to show randomness of the dynamic system. They are generated from a normal distribution with 0 mean, and $\sigma_F^2 \cdot dt, \sigma_S^2 \cdot dt$ to be the variance respectively. Whenever the density of algae equals zero, we reset the density of alga back to $10^8 \text{ cells}/L^{-1}$ as a way to ensure the adequacy of food.

Using the Lotka-Volterra model with enough food supply, we investigate how the population dynamics of *Daphnia* are influenced by the availability of sufficient algae. If this model provides a better description of the observed data than the Logistic model, it is reasonable to conclude that the density of algae plays a significant role in *Daphnia* population dynamics. On the other hand, if the model does not fit the data well, it's possible that the native algae dynamics alone may not be an adequate description of the data. Therefore, because some food management was conducted in each unit during the experiment, it may be necessary to test the Lotka-Volterra model with food supply management, as is described in the next model.

4.3 Lotka-Volterra model with food supply management

In the third model, we denote the density of alga as F (10^5 cells per liter), and density of *Daphnia* as S (individuals per liter), and total amount of *Daphnia* that is sampled out from the bucket as M (individuals per liter). The equations that describe the dynamics for this model are given below:

$$dF(t) = \alpha F(t) \left(1 - \frac{F(t)}{k_f}\right) \cdot dt + F(t) \cdot d\zeta_F - \beta F(t) S(t) \cdot dt - \delta F(t) \cdot dt + \mu \cdot dt \quad (12)$$

$$dS(t) = \theta S(t) F(t) \cdot dt + S(t) d\zeta_S - \gamma S(t) \cdot dt - \delta S(t) \cdot dt \quad (13)$$

$$d\zeta_F \sim N(0, \sigma_F^2 \cdot dt) \quad (14)$$

$$d\zeta_S \sim N(0, \sigma_S^2 \cdot dt) \quad (15)$$

$$dM = \delta S(t) \cdot dt \quad (16)$$

In this model, we set the initial condition to be $M = 0, S = 3$ and $F = 16.667$ by Searles et al. For ease of understanding the model equations, we separate the differential equations (12)-(16) into distinct parts that each correspond to meaningful ecological phenomena:

Alga The change of the density of alga can be viewed as

$$dF_{birth}(t) = \alpha F(t) \left(1 - \frac{F(t)}{k_f} \right) dt + F(t) d\zeta_F \quad (17)$$

$$dF_{cons}(t) = \beta F(t) S(t) dt \quad (18)$$

$$dF_{spl}(t) = \delta F(t) dt \quad (19)$$

$$dF_{refill}(t) = \mu dt \quad (20)$$

$$dF(t) = dF_{birth}(t) - dF_{cons}(t) - dF_{spl}(t) + dF_{refill}(t) \quad (21)$$

In these equations, α represents the growth rate; k_f is the carrying capacity; ζ_F is the noise term; and β represents the rate at which alga are consumed. All of these are treated as unknown parameters. δ and μ are fixed parameters with $\delta = 0.013$ and $\mu = 0.37$, which model the food supply rate. We split the change of alga density (dF) into four parts: equation (17) represents the change of population due to the birth of alga and Brownian noise term; equation (18) is the amount of alga that is consumed by *Daphnia*; equations (19) and (20) indicate that change of alga density due to the process of alga refill and sampling. Together, these four equations explain the processes and causes of the changes in alga population density.

Daphnia The change of the density of *Daphnia* can be viewed as

$$dS_{grow}(t) = \theta S(t) F(t) dt + S(t) d\zeta_S \quad (22)$$

$$dS_{death}(t) = \gamma S(t) dt \quad (23)$$

$$dS_{spl}(t) = \delta S(t) dt \quad (24)$$

$$dS(t) = S_{grow}(t) - S_{death}(t) - S_{spl}(t) \quad (25)$$

In these equations, θ is the growth rate, ζ_S is the noise term, and γ is the death rate, all of which are unknown parameters. δ is a fixed parameter with $\delta = 0.013$. We split the change of *Daphnia* density (dS) into three parts. The equation (22) represents the change of population due to the birth of *Daphnia* and Brownian noise terms. While the equation (23) shows the amount of alga that is consumed by *Daphnia*, which will make the density decrease. The equation (24) indicates that change of *Daphnia* density due to the process of sampling. These three equations explain the processes and causes of the changes in *Daphnia* population density.

This model includes 9 parameters: $\gamma, \alpha, \sigma_F, \sigma_S, \theta, \beta, \delta, \mu$ and k_f . Among those parameters, α represents the rate of growth of algae. k_f shows the carrying capacity of algae in each bucket. β is the consumption rate of algae by certain species of *Daphnia*, θ shows the rate of growth of *Daphnia* with certain types of food and γ is the rate of death of certain species of *Daphnia*. Similar to the previous Lotka-Volterra model, we treat δ as a constant equal to 0.013 and $\mu = 0.37$ as a constant that implies the food refilling rate according to the experimental settings (Searle et al., 2016). ζ_F and ζ_S

are Brownian noise terms, which show the randomness of the dynamic system. They are induced by a normal distribution with 0 mean, and $\sigma_F^2 \cdot dt, \sigma_S^2 \cdot dt$ to be variance respectively. Unlike the previous model, this model restricts the total food supply, more closely mimicking the experimental setting. One explanation for the observed precipitous decline in *Daphnia* population density is that there was a meaningful decline to the food supply throughout the experiment. Therefore the model equations permit both a natural decline in food levels and an increase of *Daphnia* food corresponding to the food supply rate of the experiment. Based on the experimental manipulations, we treated the food supply rate as a constant $\mu = 0.37 \cdot 10^5$ cells per day.

Using this model, we can investigate how the population dynamics of *Daphnia* are influenced by the density of algae with the condition that food levels without any food management may not be sufficient to describe *Daphnia* population dynamics. If the data are better explained using this model than the alternative hypotheses, it can be concluded that the density of algae plays a significant role in *Daphnia* population dynamics and the lack of food is potentially the reason of the precipitous decline in *Daphnia* population, as other factors may not obviously affect this treatment. Alternatively, if the model does not fit the data well, it's possible that the alga may not be the main reason that led to the change of population of *Daphnia*, which means other possible factors that may affect the population density of *Daphnia* should be considered.

4.4 Measurement and Initial condition

Important to the POMP framework is the description of how observations on the system are obtained, often referred to as the measurement model. The sampling process illustrates the potential for a difference between the true population density of *Daphnia* and that observed in the experiment due to partial sampling. As research group (Searle et al., 2016) described, on every sampling occasion, one liter of specimen is extracted from a 15 liter mixture post agitation, and the sampling rate δ is set to 0.013. Due to the constraint that both *Daphnia* and algae densities must be non-negative values, the sampling procedure was assumed to be a binomial distribution with the probability of success (p) equivalent to the sampling rate (δ) and the total count (n) as S/δ for both logistic growth models and Lotka-Volterra models. For logistic growth models, the assessment framework can be summarized as:

$$F \sim \text{binomial}(n = S/\delta, p = \delta), \delta = 0.013 \quad (26)$$

S represents the actual population concentration, and the integer $(1/\delta * S)$ signifies the estimated count of adult *Daphnia* in the 15L medium. F symbolizes the unobservable specimen resulting from the hidden population density. Given that algae samples are not accounted for in the Lotka-Volterra POMP model, we propose two different measurement models for the *Daphnia* population. The first is the same measurement model as the logistic growth model. Alternatively, we can view the process of sampling as a Poisson process, where the rate $\lambda = S \cdot \delta$ represents the average number of *Daphnia* per liter per sampling time. Hence, two possible assessment

frameworks are:

$$F \sim \text{binomial}(n = S/\delta, p = \delta), \delta = 0.013 \quad (27)$$

$$F \sim \text{Poisson}(\lambda = S \times \delta) \quad (28)$$

5 Results

In the PanelPOMP framework, one must decide whether or not each parameter should be unit specific, or shared across units. These decisions greatly affect the number of parameters fitted for each model. For example, with ten units in the model, adding a shared parameter increases the model’s dimension by one, and adding a unit specific parameter will increase the dimension by ten. An increase in parameter dimension increases the model’s ability to quantitatively describe the observed data, but leads to the possibility of over-fitting, resulting in a model with poor explanatory and predictive power. In order to find a balance between fitting accuracy and simplicity when selecting models, we employed the Akaike information criterion (AIC) as a criteria to compare models while adjusting for dimension. We estimate model parameters via maximum likelihood using the panel iterated filtering (Bretó et al., 2020) and calculated AIC for all models. Table 5 suggests that the Lotka-Volterra model that treats σ_S as a unit specific parameter and all other parameters shared results in the best AIC of 970.62. The lowest AIC of all of the considered models was the logistic growth model with only k as a unit specific parameter; this model had an AIC value of 2543.04. These results suggest that the new Lotka-Volterra model outperforms both benchmark models in fitting *Daphnia* panel data, with an AIC advantage of 1292.89 units over the best alternative model. This finding implies that the hidden alga process influences *Daphnia* population dynamics. *Daphnia* population density fluctuations primarily stem from the interplay between algae and *Daphnia*. Additionally, significant noise is generated in the *Daphnia* population throughout the process within these containers.

The results of independent likelihood maximization searches are included in the appendix. These results demonstrate that among the 250 independent searches, the vast majority of them are approaching a certain upper bound. Due to the theoretical properties of the PIF algorithm that suggest that any single iteration should approach a region near the MLE with sufficient number of particles and iterations, this is suggestive that the independent replicates are in fact converging near the MLE. Therefore we believe that we have conducted a sufficient search to find a set of model parameters that maximize the likelihood. Furthermore, the mean value of standard error of the estimated log-likelihoods is 0.114, with over 90 percent having standard error less than 1, which is small enough to provide additional evidence that proper optimization of model parameters has been achieved.

Parameter estimates and their units are shown in Tables 5, 6, and 7 in the appendix. σ_S vary across the panel while other parameters are constant across units. σ_S is the main constraints of *Daphnia* population dynamics, where σ_S is the noise generated in the population of *Daphnia*. The noise can potentially be generated due to the sampling process. Their variation in different units is possibly caused by the different

density of algae's food: higher σ_S indicates one unit has significant noise during the process. This hypothesis is plausible because the level of noise in an experiment is difficult to control in practice. By affecting the sampling process of *Daphnia*, the researchers indirectly change the data we used to interpret the performance of *Daphnia*, resulting in different dynamics across the panel.

5.1 Simulations

For the model and parameters swarm that result in the best AIC, we ran 20 simulations on each unit of the model (Figure 3 and 4). These figures demonstrate that the majority of the simulations closely resemble the observed data, suggesting the observations are a plausible realization of the fitted model. The estimated parameters that were used for this simulation are given in Tables 1 and 2.

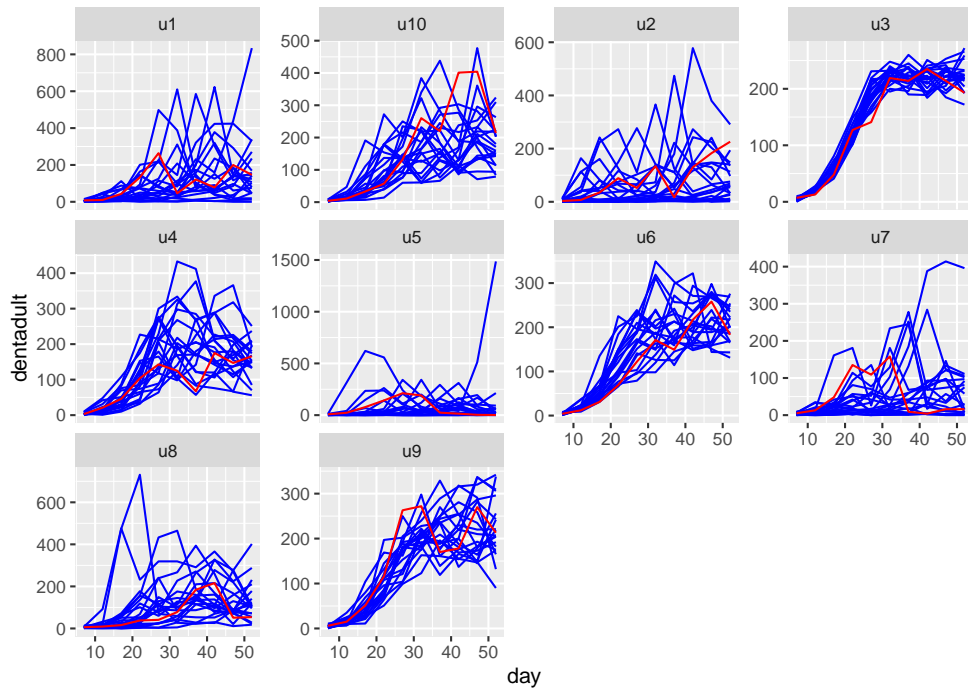


Figure 3: Simulations of the Lotka-Volterra model with σ_S to be unit specific and all the other parameters to be shared on each unit of *Daphnia dentifera* panel data.

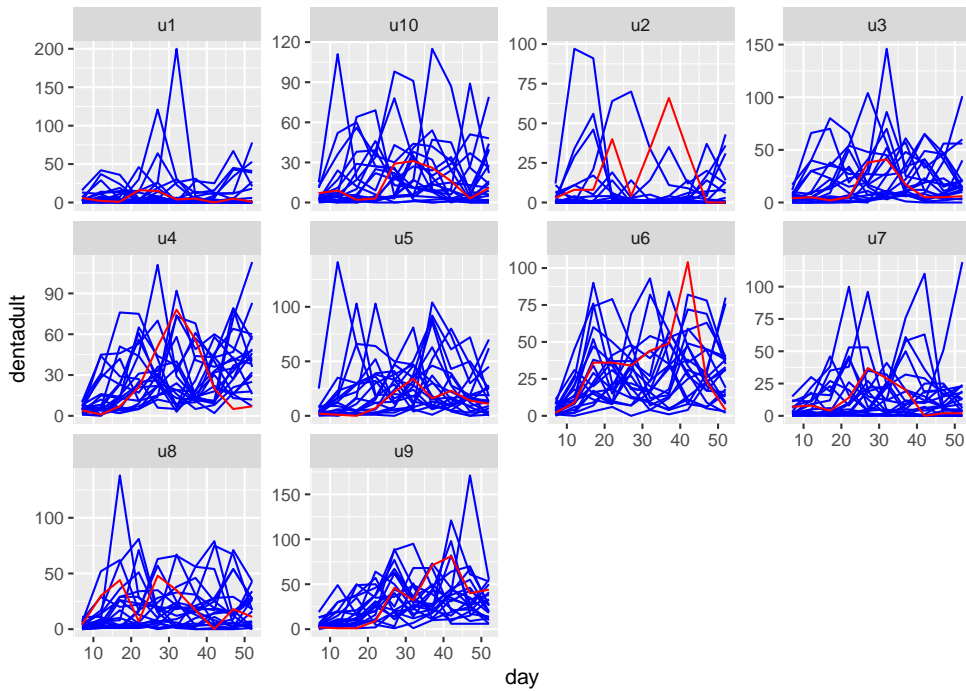


Figure 4: Simulations of the Lotka-Volterra model with σ_S to be unit specific and all the other parameters to be shared on each unit of *Daphnia lumholtzi* panel data.

6 Discussion

Seales’s research group (Searle et al., 2016) utilized a logistic growth model to depict the population dynamics of *Daphnia* in the absence of parasites. However, our findings suggest that the density of *Daphnia* is influenced to some extent by algae, which affects its growth rate. Consequently, it appears that the fluctuations in the density of *Daphnia* can be attributed to the dynamics of algae, even in the presence of parasites. Further investigation through the application of a Lotka-Volterra model on single-species with parasite treatments and mixed-species treatments may provide additional insights into the impact of algae and parasites on *Daphnia*, given enough time.

There is a concern with the model fitting since, during simulations, the density of *Daphnia* occasionally exhibits exponential growth. This may be due to the lack of a parameter restricting the upper limit of *Daphnia*’s density, which is a potential avenue to enhance the model’s accuracy in future iterations. However, I chose not to incorporate this parameter into the current model as it maintains both simplicity and relevance. Results indicate that σ_S could be the probable source of disparities in data among buckets. As this parameter is not involved in the density change of algae, it is plausible that this noise arises from some ecological process specific to *Daphnia*, which requires further scrutiny and discussion in future research.

We have demonstrated the applicability of a panelPOMP model for mechanistic modeling in ecological studies using *Daphnia* panel data. We used panelPOMP to successfully explain the large relationship between *Daphnia* population density and its food in Searle et al.’s experiment. It is worth noting that we only had 100 data

points to explain this phenomenon. In the framework of panelPOMP, we used PIF to reduce the dimensionality of the model quickly, which means that this method can be utilized for the analysis of other high-dimensional data and has the advantage of rapid dimensionality reduction. This approach complements the disadvantages of the Mechanistic model and potentially allows ecologists to try to solve more data analysis problems using the Mechanistic model.

The R package panelPOMP, which is based on this model and PIF algorithm (Bretó et al., 2020), enables access to latent state variables such as algae, offering a quantitative approach that facilitates parameter estimation, hypothesis testing, and the comparison of different models with or without their influence on the dynamic system. This approach provides scientists with better insight into the contribution of each factor to the whole system and the ability to predict its future state. Furthermore, the package allows for simulations, which serve as a valuable tool for testing the goodness-of-fit of chosen models and parameters. Overall, mechanistic modeling under a panelPOMP framework can yield more information from limited ecological data, increasing the efficiency and accuracy of ecological experiments while reducing cost and resource requirements.

7 Acknowledgement

I would like to thank Dr. Edward Ionides of the University of Michigan's Statistics Department for his exceptional assistance in crafting this dissertation. His guidance and specialized insight were invaluable to the research. Additionally, I am grateful to Jesse Wheeler, also from the University of Michigan's Statistics Department, who familiarized me with the panelPOMP framework and answered my questions about it with great patience. Furthermore, I express my appreciation to Professor Aaron King for acquainting me with the knowledge of *Daphnia* ecology and models in complex systems.

8 Appendix

Tables for each simulations and calculations

Specific parameters	dimension	max log-likelihood	AIC
K	12	-1259.52	2543.04
$K + \sigma_S$	21	-1257.07	2556.13
all specific	3	-1328.33	2662.67
σ_S	12	-1366.01	2756.03
$r + K$	21	-2011.21	4065.42
$r + \sigma_S$	21	-2111.21	4264.42
r	12	-2138.59	4301.19

Table 1: This table includes logistic growth models fitted to *Daphnia dentifera* with different choices of parameters to be either unit specific or shared

Specific parameters	dimension	max log-likelihood	AIC
α	16	-1122.15	2276.30
β	16	-1166.73	2366.80
γ	16	-1167.40	2365.46
θ	16	-1176.68	2385.35
σ_S	16	-1249.57	2531.15
σ_F	16	-1242.07	2516.14
k_f	16	-1157.30	2346.60
α, k_f	25	-1106.76	2263.51
α, β, k_f	34	-1103.92	2275.85
β, k_f	25	-1114.36	2278.73
α, γ, k_f	34	-1112.00	2292.00
θ, k_f	25	-1131.60	2313.19
α, θ	25	-1131.60	2313.19
all specific	70	-1099.39	2338.78

Table 2: This table includes Lotka-Volterra model with enough food supply fitted to *Daphnia dentifera* with different choices of parameters to be either unit specific or shared

Specific parameters	dimension	max log-likelihood	AIC
σ_S	16	-469.31	970.62
σ_S, θ	25	-461.31	972.62
σ_S, β	25	-463.32	976.64
σ_S, α	25	-463.48	976.96
σ_S, γ	25	-465.65	981.31
σ_S, σ_F	25	-466.95	983.9
σ_S, k_f	25	-465.71	980.44
α, σ_F	25	-483.72	1017.41
α, β	25	-493.94	1037.88
α	16	-501.31	1034.62
β, γ	25	-503.31	1056.63
γ, σ_F	25	-503.51	1057.03
β, σ_F	25	-503.74	1057.48
σ_F	16	-513.73	1059.46
α, θ	25	-505.64	1061.31
θ, γ	25	-506.72	1063.45
α, γ	25	-506.81	1063.61
β	16	-516.31	1064.62
k_f, σ_F	25	-507.77	1065.55
θ, β	25	-511.95	1073.91
θ	16	-521.31	1074.62
θ, β	25	-511.95	1073.91
α, θ	25	-512.55	1075.10
k_f	16	-523.73	1079.46
α, k_f	25	-515.36	1080.72
γ, k_f	25	-515.36	1080.73
θ, k_f	25	-519.21	1088.47
β, k_f	25	-520.04	1090.09

Table 3: This table includes the Lotka-Volterra models with food management fitted to *Daphnia dentifra* with different choices of parameters to be either unit specific or shared

Specific parameters	dimension	max log-likelihood	AIC
σ_S	16	-368.18	768.38
β	16	-369.17	770.34
γ	16	-369.75	771.51
α, β	25	-361.04	772.08
α	16	-370.37	772.74
σ_S, β	25	-362.59	775.18
β, σ_F	25	-362.99	775.98
θ	16	-372.28	776.56
k_f	16	-372.37	776.74
σ_S, α	25	-363.71	777.42
σ_F	16	-373.28	778.56
α, σ_F	25	-365.35	780.71
σ_S, σ_F	25	-366.04	782.08
σ_S, k_f	25	-366.56	783.12
σ_S, θ	25	-366.67	783.34
σ_S, γ	25	-366.81	783.62
θ, β	25	-369.31	788.62
α, θ	25	-370.24	790.48
α, k_f	25	-371.13	792.26
α, γ	25	-371.15	792.31
k_f, σ_F	25	-372.81	795.62
θ, σ_F	25	-373.11	796.22
γ, σ_F	25	-372.57	795.14
β, γ	25	-373.19	796.38
β, k_f	25	-373.48	796.96
θ, γ	25	-373.75	797.51
θ, k_f	25	-374.37	798.74
γ, k_f	25	-375.56	801.12

Table 4: This table includes the Lotka-Volterra models with food management fitted to *Daphnia lumzniza* with different choices of parameters to be either unit specific or shared



Figure 5: This figure describes the change of the corresponding loglikelihood in the process of iteration under different particles, we only chose the graph after the twentieth iteration, in order to be able to see more clearly that they are close to the upper bound.

unit	$\alpha \cdot 10$	$\beta \cdot 10^4$	$\gamma \cdot 10$	$\theta \cdot 10$	k_f	$\sigma_S \cdot 10$	$\sigma_F \cdot 10$
u1	4.32	8.15	3.61	4.08	0.25	4.3	0.09
u2	4.32	8.15	3.61	4.08	0.25	5.31	0.09
u3	4.32	8.15	3.61	4.08	0.25	0.39	0.09
u4	4.32	8.15	3.61	4.08	0.25	2.37	0.09
u5	4.32	8.15	3.61	4.08	0.25	5.49	0.09
u6	4.32	8.15	3.61	4.08	0.25	1.37	0.09
u7	4.32	8.15	3.61	4.08	0.25	5.39	0.09
u8	4.32	8.15	3.61	4.08	0.25	3.9	0.09
u9	4.32	8.15	3.61	4.08	0.25	1.38	0.09
u10	4.32	8.15	3.61	4.08	0.25	2.28	0.09

Table 5: This table includes parameter swarm that gets best AIC with Lotka-Volterra models fitted on *Daphnia dentifera* panel data

unit	$\alpha \cdot 10$	$\beta \cdot 10^4$	γ	$\theta \cdot 10$	k_f	$\sigma_S \cdot 10$	$\sigma_F \cdot 10$
u1	4.44	2.43	9.58	3.31	3.25	5.8	2.36
u2	4.44	2.43	9.58	3.31	3.25	6.55	2.36
u3	4.44	2.43	9.58	3.31	3.25	3.16	2.36
u4	4.44	2.43	9.58	3.31	3.25	2.04	2.36
u5	4.44	2.43	9.58	3.31	3.25	3.14	2.36
u6	4.44	2.43	9.58	3.31	3.25	1.6	2.36
u7	4.44	2.43	9.58	3.31	3.25	4.66	2.36
u8	4.44	2.43	9.58	3.31	3.25	4.54	2.36
u9	4.44	2.43	9.58	3.31	3.25	1.85	2.36
u10	4.44	2.43	9.58	3.31	3.25	3.9	2.36

Table 6: This table includes parameter swarm that gets best AIC with Lotka-Volterra models fitted on *Daphnia lumholtzi* panel data.

parameter	unit
α	day^{-1}
β	$L \cdot individual^{-1} \cdot day^{-1}$
γ	$(6hr)^{-1}$
θ	$L \cdot (10^5 cells)^{-1} \cdot day^{-1}$
k_f	$10^5 cells \cdot L$
σ_S	day^{-1}
σ_F	day^{-1}

Table 7: This table shows the units of the parameters

References

- Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-gaussian bayesian tracking. *TSP*, 50:174 – 188.
- Baer, K. N. and Goulden, C. E. (1998). Evaluation of a high-hardness combo medium and frozen algae for daphnia magna. *Ecotoxicology and environmental safety*, 39:201–306.
- Bretó, C., Ionides, E. L., and King, A. A. (2020). Panel data analysis via mechanistic models. *Journal of the American Statistical Association*, 115(531):1178–1188.
- Duarte, C. M., Amthor, J. S., DeAngelis, D. L., Joyce, L. A., Maranger, R. J., Pace, M. L., Pastor, J., and Running, S. W. (2003). *The limits to models in ecology*, pages 437–451. Princeton University Press, New Jersey, NY.
- Holme, P. and Liljeros, F. (2015). Mechanistic models in computational social science. *Frontiers in Physics*, 3.
- Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724.
- King, A. A., Nguyen, D., and Ionides, E. L. (2016). Statistical inference for partially observed markov processes via the R package pomp. *Journal of Statistical Software*, 69(12).
- Lomnicki, A. (1980). Regulation of population density due to individual differences and patchy environment. *Oikos*, 35(2):185–193.
- Searle, C. L., Cortez, M. H., Hunsberger, K. K., Grippi, D. C., Oleksy, I. A., Shaw, C. L., de la Serna, S. B., Lash, C. L., Dhir, K. L., and Duffy, M. A. (2016). Population density, not host competence, drives patterns of disease in an invaded community. *The American Naturalist*, 188(5):554–566.
- Stalidzans, E., Zanin, M., Tieri, P., Castiglione, F., Polster, A., Scheiner, S., and Schmidt, H. H. (2020). Mechanistic modeling and multiscale applications for precision medicine: Theory and practice. *Network and systems medicine*, 3(1):36–56.
- Yang, X. (2018). Fitting mechanistic models to daphnia panel data within a panelpomp framework. Honor thesis, Department of Statistics, University of Michigan. https://ionides.github.io/students/xyang_honors_thesis.pdf.