Fitting mechanistic models to *Daphnia* panel data within a panelPOMP framework

Xiaotong Yang

Supervisor: Prof. Dr. Edward Ionides

Secondary supervisor: Dr. Carles Bretó

May 9, 2018

An honors thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science (Honors Statistics) at the University of Michigan, 2018

Abstract

Mechanistic modeling in ecological studies is always an intriguing but challenging topic. On one hand, mechanistic models allow quantitative analysis, provide better insight of the dynamic system and enable inference and prediction. On the other hand, however, it's constrained by the nonlinear, non-Gaussian feature of most of the ecological data. Panel data is a common data type used in ecology. It contains multi-dimensional data measured over time. Analyzing panel data incurs additional challenges because of its high dimensional feature. In this study, we aim to prove the viability and advance of mechanistic modeling on ecological panel data with panelPOMP framework and the panel iterated filtering maximization algorithm (PIF). We fit logistic growth model and predator-prey model on *Daphnia* panel data collected by Searle et al. (2016) with 10 independent time series, each including 10 data points. We perform PIF as likelihood maximization and parameter estimation method and compare two models using Akaike information criterion (AIC). The result shows that predator-prey model has better AIC, which indicates prey of *Daphnia* has a significant influence on *Daphnia* population dynamics. With this example, we illustrate how a panelPOMP model overcomes the nonlinearity and high dimensional structure in panel data analysis, highlight its explanatory power on the latent process with no associated observable data and emphasize quantitative insight it grants.

1 Introduction

Mechanistic models, also known as hierarchical or multilevel explanatory models, involves multiple biological organizations, where one is observable while other organizations are unobservable. With the known process, they explain the behavior of latent dynamics (Duarte et al., 2003). Mechanistic modeling in ecology aims at describing the evolvement and relationship within a dynamic system with mathematical models. As a quantitative tool, it allows scientists to compare the explanatory powers of different theoretical models and estimate the magnitude of parameters of interest. Mechanistic modeling has a long history in scientific studies and has been widely used in social sciences, industrial engineering, and physics. Nowadays, it has become more efficient with the development of computing power. However, because of the non-stationary, nonlinear and stochastic features of ecological data, traditional time series analysis has limited explanatory power over it. Even though scientists have theoretical models for their data, how to bridge models to data remains challenging. As a result, mechanistic modeling only plays a limited role in ecological research (Duarte et al., 2003).

Panel data or longitudinal data refers to a collection of independent outcomes under certain treatment or exposures at multiple times points. As a common data type in ecology, scientists collect panel data as a way to minimize the effect of randomness, measure the change of individual unit and control the outside effect. The analysis of panel data not only involves the nonlinear, non-Gaussian properties of ecological data but also a high-dimensional structure. All the features weaken the explanatory power of traditional time series analysis as well as some Monte Carlo inference approaches. In this study, we aim to apply mechanistic models on ecological panel data within panelPOMP framework and use panel iterated filtering (PIF) as maximization algorithm (Bretó et al., 2018). panelPOMP is an adaptation of partially observable Markov Process model on panel data by constructing a POMP model on each unit. panelPOMP allows testing whether each parameter in the chosen mathematical model is shared or specific across the panel by calculating the maximum log likelihood of all possible combinations with panel iterated filtering algorithm. Panel iterated filtering derives from iterated filtering (Ionides et al., 2006) on single nonlinear stochastic time series data. Besides filtering within each time series, it also cycles through the panel to reach the highest likelihood, which both allows to apply mechanistic models on panel data and reserve its high dimensional feature.

To better illustrate the advance of panelPOMP model, we constructed two traditional ecological models: logistic growth model and predator-prey model in panelPOMP framework and fitted them on *Daphnia* panel data provided by Searle et al (2016). In their study, "Population Density, Not Host Competence, Drives Patterns of Disease in an Invaded Community", the authors explored how host densities influenced the disease pattern by studying how one invasive*Daphnia* species affects one native *Daphnia* and its parasite. Their lab experiment produced a typical ecological panel data, including multiple independent replications under the same treatment, each of which contains a short nonlinear, non-stationary time series. Traditional mechanistic models fail to directly explain these data. With the help of the panelPomp framework, we can combine theoretical models with empirical data and extract the full information contained in each time series. After fitting both models on *Daphnia* panel data, we calculated their maximum log-likelihood and estimated the optimal parameter swarm via iterated filtering algorithm. Results showed the predator-prey model has better performance, which indicates *Daphnia*'s prey, whose population density is unobservable in data, also influences daphnia population dynamics. In section 2.1 we will describe the *Daphnia* dataset in detail; in section 2.2–2.4 we will explain the panelPOMP model and the model setup. Section 2.5 will explain the inference methodology of panelPOMP: panel iterated filtering algorithm in detail. In section 3, we will present and explain the result of this example. The last section will include the significance of panelPOMP method in mechanistic modeling and some future improvement for this example.

With this example, we hope to compare the explanatory power of two classic ecological models on the *Daphnia*panel data, estimate the key parameters that help understand the evolution of the ecological system, demonstrate the viability and advance of mechanistic models in the ecological study and more importantly, how panelPOMP models and PIF help overcome the internal difficulties of ecological model fitting.

2 Methodology

2.1 Data description

Searle et al. (2016) first performed individual-level experiment to quantify the host competence of a native and an invasive host species. Then they designed a mesocosm experiment to capture community-level species interaction and used a mathematical model to explain the role population density played in disease patterns. They chose *Daphnia* dentifera as native species, *Daphnia* lumholtzi as invasive species and fungus Metschnikowia bicuspidate as parasites of this community. The experiment was conducted in indoor mesocosm, with 6 different treatments: native species only with and without parasite, invasive species only with and without parasite and both species combination with and without parasites. 45 *Daphnia* were added into 15L COMBO media at initiation in all four single-species treatments. 35 native and 10 invasive species were added in mix-species treatments. Each treatment was replicated 10 times. This experiment lasted for 52 days and sampled every five days starting seventh days after experiment started. On each sampling day, 1L out of total 15L solution is removed as sample for each of the 10 replicates after vigorously stirring and replaced by 1L of fresh COMBO media. Species, infection status, age, and sex were recorded for each individual in the removed sample. Temperature in the lab was kept constant and media solution in different replicates remains the same over time.

In this study, we focused only on the native species, *Daphnia dentifera*, when the parasite is absent. This panel data contains 10 independent time series of dentifera adult population density, all collected at 10 same time points (figure 1). These data allow us to compare mechanistic models and test the impact of their food, algae, on *Daphnia* population dynamics.

2.2 The panelPOMP model

A panel data with u independent time series has u units denoted as $1, 2, \ldots, U$, where each unit has n measurement collected at times $(t_1, t_2, \ldots, t_{N_u})$. These measurement $y^*_{u,1}, \ldots, y^*_{u,N_u}$ is generated by a stochastic observable process $Y_{u,1}, \ldots, y_{u,N_u}$. PanelPOMP is an adaptation of partially observable Markov process (POMP) (Bretó et al., 2018) on panel data analysis, involving constructing POMP model on each unit and find the maximum likelihood with panel iterated filtering algorithm. For each time series unit, we built a process model and a measurement model as the skeleton of POMP framework(King et al., 2016). Process model describes the real population dynamics with the one-step transition density$f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$ and measurement model explains how the latent real population density generates the observable time series data with the measurement density $f_{Y_n|X_n}(y_n|x_n;\theta)$. In summary, the joint density will be the product of the process model and the measurement model at each state times the initial density $f_{X_0}(x_0;\theta)$. In this study, we want to fit and compare two traditional ecological models, a logistic growth model and a predator-prey model under panelPOMP framework.

2.3 Process model

Process models show how the states of the latent process are updated and how the dynamic systems evolve. The process model of logistic growth model is represented by the following stochastic equation:

$$\frac{dS}{dt} = (r + noiG)S(1 - \frac{S}{K}) - \delta S$$

This model contains one state variable *Daphnia* population density(S) and 3 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. δ denotes the sampling rate in the experiment. Out of dimension reduction concern, we fixed δ as a constant equal 0.013 as in Searles et al. (2016). $noiG$ is a noise term added to population growth to allows for random variation in the dynamic. In this case, we assure $noiG$ follows a Normal distribution with 0 mean and a standard deviation of σ_G.

$$noiG \sim N(0, \sigma_G)$$

This classic ecological model is designed to capture the change of growth speed of *Daphnia* population in a relatively close lab environment, where maximum capacity restricts the highest possible density. As population density(S) approach K the growth speed of population size keeps decreasing and reaches 0 when S equal to K. The population will shrink if S exceeds K.

The alternative model aims at describing the relationship between algae and *Daphnia* population density with the predator-prey model. This model has two state variables: algae density(F) and *Daphnia* density(S), where *Daphnia* population generated an observable measurement but algae didn't. The stochastic equations are as follows:

$$\frac{dF}{dt} = (\alpha + noiB)F(1 - \frac{F}{k_f}) - \beta SF$$

$$\frac{dS}{dt} = (\theta + noiG)SF - \gamma S - \delta S$$

This model includes 7 parameters: α, k_f, β, γ, θ, σ_B, and σ_G. Parameter α denotes the growth rate of algae. k_f is the carrying capacity of algae. β represents the intake rate of algae by *Daphnia*, θ is the growth rate of *Daphnia* given algae and γ is the *Daphnia*'s death rate. Same as the logistic model, we treat δ as a constant equal to 0.013 to keep the simplicity of the model. $noiB$ and $noiG$ are noise terms added to algae (F) growth and *Daphnia* (S) growth to induce random variation to this system. Both of them are generated by a normal distribution with 0 mean,

$$noiB \sim N(0, \sigma_B)$$

$$noiG \sim N(0, \sigma_G)$$

This model allows us to test the impact of algae on *Daphnia* population dynamics. Since no other factors such as invasive species, parasite or low temperature can affect this treatment, if the alternative explains the data well, we can conclude that algae play an important role in *Daphnia* population dynamics.

2.4 Measurement model and initiation of dynamic

The measurement model demonstrates the process that latent real population density generates observable data. In this case, it involves drawing samples from the *Daphnia* population distribution. According to Searles et al. (2016) , on each sampling day, 1L of sample is removed from 15L solution after stirring and the sampling rate is 0.013. Since the number of *Daphnia* and algae is restricted to be positive integers, we treated the process as binomial distributions with success probability (p) equal to sampling rate (δ) for both the logistic growth and the predator-prey panelPOMP models. In the logistic model, the measurement model is:

$$F \sim binomial(n = nearbyint(1/\delta * S), p = \delta)$$

Where S is the real population density and $nearbyint(1/\delta * S)$ is the approximate real quantity of *Daphnia* adult in 15L media. F denotes the observable sample generated from latent population density. Since algae have no observable sample from the data, in predator-prey pomp object, we only constructed the measurement model for *Daphnia* population, which is the same as that of the logistic model.

$$F \sim binomial(n = nearbyint(1/\delta * S), p = \delta)$$

2.5 Panel iterated filtering (PIF)

Panel iterated filtering (PIF) derives from IF2 algorithm for POMP likelihood maximization (Ionides et al., 2015). PIF treats each unit in a panelPOMP model as a time-inhomogeneous POMP model. Given a POMP skeleton and the initial guess of parameters, the algorithm cycles through the panel from the first unit to the last in each iteration to reach the maximum likelihood. That means the initial parameter swarm of unit u is simulated with the best parameter swarm of unit $u-1$. Since PIF allows every parameter either share the same value across the panel or achieve different value in each unit, each model has 2^n different combination during parameter estimation, where n denotes the number of parameters. To estimate the best parameter values, we tested all possible situations. We performed 10 replications of PIF with 250 iterations and 10,000 particles for each situation, chose the best log likelihood in those replications and recorded its optimal parameter swarm for that situation. All the code can be found in online repository https://bitbucket.org/xiaotongyang/daphnia-panelpomp.

3 Results

Allowing parameters to be shared or specific incur different dimension of each model. Since the panel has 10 units, one more shared parameter increases the dimension of the model by 1 while one specific parameter incurs 10 more dimensions. We faced the trade-off between goodness of fit and simplicity during model selection. Therefore, we used Akaike information criterion (AIC) to compare models with adjustment of dimension. We estimated parameters and maximum log-likelihood with panel iterated filtering and calculated AIC for both models (Table1&2). As shown in tables, the predator-prey model with α, k_f specific and all other parameters shared yield the best AIC 2263.51. The logistic growth model with k specific and all other shared has the smallest AIC 2543.04. We can safely conclude predator-prey model has better performance than logistic growth model on fitting *Daphnia* panel data: AIC of the former is 279.53 units better than that of the later. This result suggests latent process algae does affect the population dynamic of *Daphnia*. The fluctuation of *Daphnia* population density is mainly caused by the interaction between algae and *Daphnia*.

Parameter estimates and confidence intervals are shown in table 3. α and k_f vary across the panel while other parameters are constant across units . α and k_f are two main constraints of algae population dynamics, where α describe its growth rate and k_f is the maximum capacity. Their variation in different units is probably caused by the different density of algae's food: higher α and k_f indicates one unit has more adequate food for algae than others. Level of algae's food is difficult to control in practice. By affecting the density of algae, they indirectly change the performance of *Daphnia*, resulting in different dynamics across the panel.

For the best model fit, we plot 20 simulations on each time series in this panel data using the parameters estimated by PIF (Figure 2). From the figure, all but three simulations capture the data well.

4 Discussion

In their study, Seales et al. (2016) described *Daphnia* population dynamic without parasite using logistic growth model. However, our result indicates that algae at least partially influence the density of *Daphnia* by affecting its growth rate. This suggests the dynamic of algae is responsible for the fluctuation of *Daphnia* density even when the parasite is incurred in the system. With more time, we could apply the predator-prey model to single species with parasite treatments and mixed species treatments to test how algae and parasite affect *Daphnia*.

One concern about the model fitting is that simulations on three units in the panel failed to completely capture the data: unit 5, unit 7, and unit 8 (figure 2). Their time series all contain an extreme peak. This phenomenon might results in the measurement error during sampling. Since samples are collected every 5 days and each time series only has 10 data points, with one or two outliers, the trend of the whole time series could be changed. Another possible explanation is the algae density also contributes to the death rate of *Daphnia*. In the predator-prey model, the prey affects the predator only through the growth rate θ and death rate γ of predator is independent of algae dynamics. When the prey density is low, the growth speed of predator decreases or even turns negative but predator's death rate won't be affected by starvation. In this panel, all three units involve sharp drops in *Daphnia* density, but the change in growth rate itself failed to perform the decrease fast enough. In this case, the reduction can be caused by a combination of lower growth rate and higher death rate, which exceeds the explanatory power of the predator-prey model. Limited by time, the modification of death term in this model remains as future work.

With *Daphnia* panel data, we demonstrated how a panelPOMP model allows mechanistic modeling in ecological studies. R package panelPomp has been developed based on panelPOMP model and PIF algorithm (Bretó et al., 2018). Compared with traditional qualitative analysis, panelPomp package provides us access to latent state variables such as algae. We can test different models with or without their influence on the dynamic system by comparing maximum log likelihoods calculated with PIF. The quantitative approach also enables parameter estimation and hypothesis testing. With the knowledge of the direction and magnitude of each parameter of interest, scientists will have better insight on the contribution of each factor on the whole system and predict its evolution in the future state. The package also makes simulation possible, which works as another powerful tool for scientists to test the goodness-of-fit of their chosen models and parameters.In summary, gathering ecological data can be difficult and costly, mechanistic modeling under a panelPOMP framework digs more information from limited data, which can largely increase the efficiency of ecological experiments.

5 Acknowledgment

I would like to thank *Dr. Edward Ionides*, from the *Department of Statistics* at the *University of Michigan*, for his extraordinary support in writing this thesis. He provided the general direction and expert advice to this study. I also receive help from *Dr. Carles Bretó* from the *Department of Statistics* at the *University of Michigan.* He introduced me to the panelPomp package and answered my questions about it with great patience. In addition, I would like to thank *Camden Gowler* from the *Duffy lab* for introducing us to *Daphnia* ecology.

References

[1] Bretó, C., Ionides, E. L., King, A. A. (2018). Panel data analysis via mechanistic models. Retrieved from https://arxiv.org/abs/1801.05695.

[2] Duarte, C. M., Amthor, J. S., DeAngelis, D. L., Joyce, L. A., Maranger, R. J., Pace, M. L., Running, S. W. (2003). The Limits to Models in Ecology. In Models in Ecosystem Science (pp. 437-451). New Jersey, NY: Princeton University Press.

[3] Ionides, E. L., Nguyen, D., Atchade, 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 of the USA 112 719-724.

[4] King, A. A., Nguyen, D., and Ionides, E. L.2016. Statistical Inference for Partially Observed Markov Processes via theRPackagepomp. Journal of Statistical Software, 69(12). doi:10.18637/jss.v069.i12.

[5] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[6] Searle, C. L., Cortez, M. H., Hunsberger, K. K., Grippi, D. C., Oleksy, I. A., Shaw, C. L., Duffy, M. A. (2016). Population Density, Not Host Competence, Drives Patterns of Disease in an Invaded Community. The American Naturalist, 188(5), 554-566. doi:10.1086/688402

Appendix

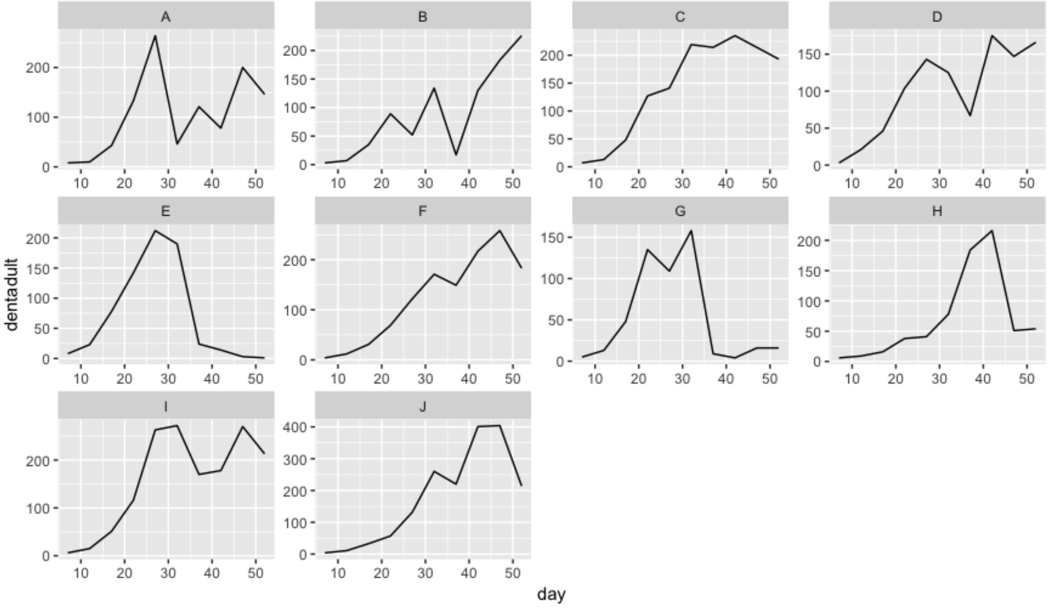

Figure 1: Adult *Daphnia* population density time series plot in native species only without parasite treatment. There are 10 replications in this treatment, denoted as A to J. Their plots are arranged from left to right, top to bottom.

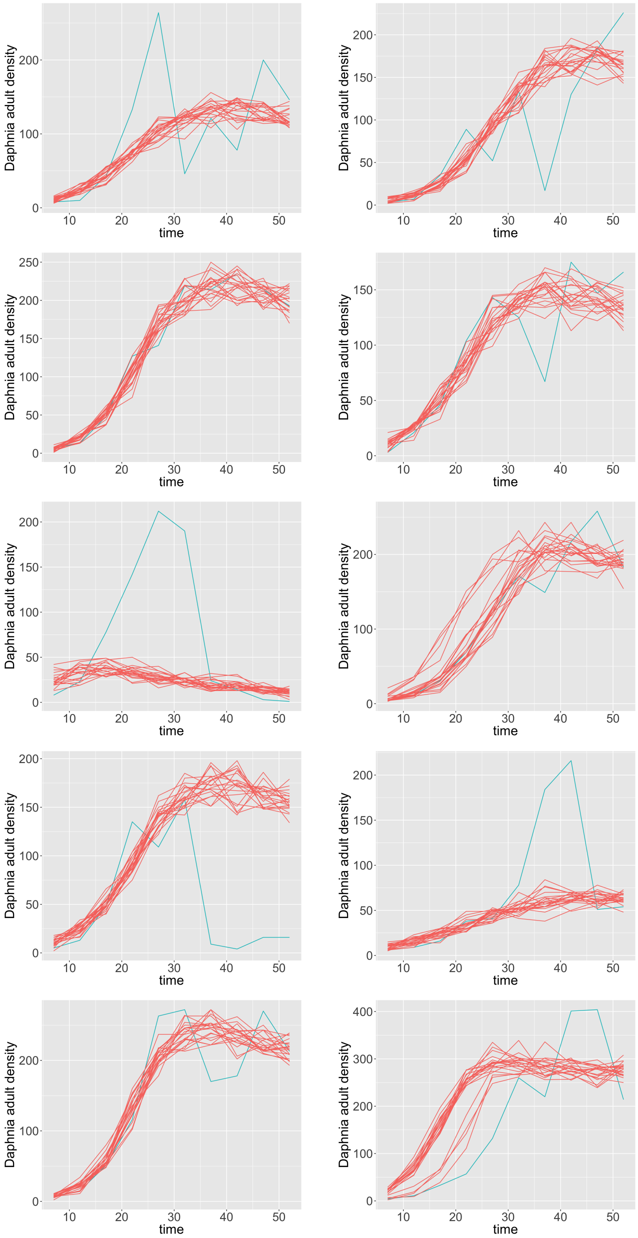

Figure 2: Simulations of the predator-prey model with α and k_f specific and the rest of parameters shared on each unit of *Daphnia* panel data. From left to right, top to bottom are unit 1, unit 2, . . . unit 10. In each plot, y-axis is *Daphnia* adult density in each unit and x-axis is time in days. The blue line represents data and red lines denotes 20 simulations with parameter swarm yield by PIF.

Specific parameters	dimension	maximum log-likelihood	AIC
$\alpha + k_f$	25	-1106.76	2263.51
$\alpha + \beta + k_f$	34	-1103.92	2275.85
α	16	-1122.15	2276.30
$\beta + k_f$	25	-1114.36	2278.73
$\alpha + \beta + \gamma + k_f$	43	-1100.04	2286.07
$\alpha + k_f + \gamma$	34	-1112.00	2292.00
$\alpha + \beta + \gamma + \theta + k_f$	52	-1094.90	2293.80
$\theta + k_f$	25	-1131.60	2313.19
$\alpha + \theta$	25	-1138.08	2326.16
all specific	70	-1099.39	2338.78
$\beta + \theta$	25	-1145.69	2341.40
k_f	16	-1157.30	2346.60
$\beta + \sigma_B$	25	-1149.37	2348.74
$k_f + \sigma_G$	25	-1157.37	2364.73
$\theta + \sigma_G$	25	-1157.51	2365.02
γ	16	-1166.73	2365.46
β	16	-1167.40	2366.80
$\gamma + k_f$	25	-1159.05	2368.10
$k_f + \sigma_B$	25	-1159.83	2369.66
$\alpha + \beta$	25	-1165.81	2381.61
θ	16	-1176.68	2385.35
$\beta + \sigma_G$	25	-1176.02	2402.04
$\gamma + \sigma_B$	25	-1181.15	2412.29
$\alpha + \sigma_G$	25	-1182.47	2414.94
$\beta + \gamma$	25	-1189.78	2429.57
$\theta + \sigma_B$	25	-1212.12	2474.24
$\gamma + \theta$	25	-1222.61	2495.22
σ_B	16	-1242.07	2516.14
$\alpha + \gamma$	25	-1234.12	2518.24
$\sigma_B + \sigma_G$	25	-1236.74	2523.47
σ_G	16	-1249.57	2531.15
$\gamma + \sigma_G$	25	-1251.52	2553.04
all shared	7	-1349.28	2712.56
$\alpha + \sigma_B$	25	-1334.27	2718.55

Table 1: Predator-prey models with different choices of parameters, dimensions, maximum log-likelihood and AIC. In each model, parameters listed in "specific parameters" are specific for each unit and the rest parameters are shared across the panel. The table is arranged by from smallest (best) AIC to largest (worst) AIC. From the table, predator-prey model with α and k_f specific and the rest of parameters shared has the best performance.

Specific parameters	dimension	maximum log-likelihood	AIC
K	12	-1259.52	2543.04
$K + \sigma_G$	21	-1257.07	2556.13
all shared	3	-1328.33	2662.67
σ_G	12	-1366.98	2757.97
$r + K$	21	-2011.87	4065.75
$r + \sigma_G$	21	-2112.44	4266.87
r	12	-2139.59	4303.19

Table 2: Logistic growth models with different choices of parameters, dimensions, maximum log-likelihood and AIC. In each model, parameters listed in "specific parameters" are specific for each unit and the rest parameters are shared across the panel. The table is arranged by from smallest (best) AIC to largest (worst) AIC. From the table, predator-prey model with K specific and the rest of parameters shared has the best performance.

unit	α	k_f	β	γ	θ	σ_B	σ_G
A	0.571	256.575	0.00388	0.0254	0.000799	0.00931	1.112e-05
B	0.758	269.797	0.00388	0.0254	0.000799	0.00931	1.112e-05
C	0.935	321.890	0.00388	0.0254	0.000799	0.00931	1.112e-05
D	0.640	273.102	0.00388	0.0254	0.000799	0.00931	1.112e-05
E	0.0158	28.650	0.00388	0.0254	0.000799	0.00931	1.112e-05
F	0.874	301.145	0.00388	0.0254	0.000799	0.00931	1.112e-05
G	0.729	287.445	0.00388	0.0254	0.000799	0.00931	1.112e-05
H	0.316	176.087	0.00388	0.0254	0.000799	0.00931	1.112e-05
I	1.017	344.067	0.00388	0.0254	0.000799	0.00931	1.112e-05
J	1.216	388.110	0.00388	0.0254	0.000799	0.00931	1.112e-05

Table 3: Parameter swarm that yields best AIC among predator-prey models, where α and k_f are specific in each unit and the rest parameters are shared.