Modeling Populations of Flour Beetles

S. Robertson

In terms of mathematical content, this case study explores the following topics: Calculating maximum likelihood parameter estimates for nonlinear stochastic models, and connecting these models with data.

1 Introduction

Mathematical models can be used to describe and predict the complex dynamics of biological phenomena such as animal populations. There is a trade-off between the amount of mathematical and biological complexity one can include in a model; fewer simplifications lead to a more biologically accurate model, but may come at the cost of adding dimensions or parameters. This increases the difficulty of mathematical analysis and may render some analyses intractable.

A critical part of evaluating a mathematical model is connecting it with data. Data is used to give numerical values to a model's parameters and to judge the adequacy of the model to describe the dynamics of a population. Once parameter values are obtained, the parameterized model can be applied to additional data to test its predictive power. Computer simulations and mathematical analysis can be used to investigate the dynamics of the model and interesting dynamics found may suggest future experiments that could further validate the model.

2 The LPA Model

Accurate population data (free from confounding variables) is difficult to find in nature, and so laboratory cultures of animals are often used to validate models. In particular, the flour beetle *Tribolium castaneum* is a very convenient and useful organism to work with in the laboratory. The flour beetle's life cycle has three stages - larva, pupa, and adult - with it taking about two weeks to move between stages. The species also exhibits nonlinear interactions between life stages, namely cannibalism of the non-moving stages by the moving stages. This allows for very interesting population dynamics.

One of the most heavily analyzed and well-validated models in mathematical ecology is a model describing flour beetle population dynamics known as the "Larvae-Pupae-Adult" (LPA) model. The model is given by the following system of three nonlinear difference equations:

$$\begin{aligned} L_{t+1} &= bA_t e^{-c_{el}L_t - c_{ea}A_t} \\ P_{t+1} &= (1-\mu_l)L_t \\ A_{t+1} &= P_t e^{-c_{pa}A_t} + (1-\mu_a)A_t \end{aligned} \tag{1}$$

L_t is the number of larvae at time t, P_t represents the number of individuals in the "P stage" (including non-feeding larvae, pupae, and callow adults) at time t, and A_t is the number of sexually mature adults at time t. The time step is two weeks, the amount of time it takes to transition to the next class. Once reaching the adult stage beetles remain adults until death. Eggs are laid by adults at a rate b, and these eggs must survive cannibalism by larvae and adults in order to become larvae. $c_{el} \geq 0$ and $c_{ea} \geq 0$ are cannibalism coefficients of eggs by larvae and eggs by adults, respectively. It is assumed that cannibalism occurs as a result of random encounters of larvae and adults with eggs, and thus exponential nonlinearities are used [1]. Larvae die at a natural death rate μ_l, $0 < \mu_l < 1$, and so a fraction $(1-\mu_l)$ survive to become pupae. The death rate of pupae is negligible, so there is no μ_p term included in the model. Pupae must escape cannibalism by adults (c_{pa}) to become adults. Adults die at a rate μ_a, $0 < \mu_a < 1$, and so at each census, the fraction of surviving adults is $(1-\mu_a)$.

(a) *Tribolium* larvae and adults.

(b) *Tribolium* cultures in R. F. Costantino's lab.

Model (1) predicts the number of larvae, pupae, and adults in a culture at time $t+1$ $(L_{t+1}, P_{t+1}, A_{t+1})$ given the number of animals in each life stage at time t (L_t, P_t, A_t), provided we have values for the parameters b, μ_l, μ_a, c_{el}, c_{ea}, and c_{pa}.

3 Parameter Estimation

3.1 Collecting Data

Data to estimate parameter values and test the validity of the LPA Model was collected by Dr. R. F. Costantino at the University of Arizona. Cultures of *T. castaneum* are started by placing an initial number of larva, pupa and adult beetles into 20 grams of flour medium. This initial population vector provides the first data point (L_0, P_0, A_0). The culture is returned to a dark incubator. Two weeks later Costantino censuses the culture again, noting the new number of larvae, pupae and adults. These numbers form the second data point (L_1, P_1, A_1). All life stages are again returned to the culture with renewed medium, censused again in two weeks, and so on. Together the observed data points form a time series (L_t, P_t, A_t), for $t = 1, ..., m$ where m is the length of the time series.

Sometimes parameter values can be estimated directly. For example, the natural adult death rate μ_a can be calculated by noting how many dead adults there are in the culture at each time step. Most parameter values are not as easily obtained and must be estimated from laboratory data using statistical techniques.

3.2 Variability in Data

If the LPA Model and its parameter values were 100% accurate, then given an observed data point (L_t, P_t, A_t) the model would always output the next observed data point $(L_{t+1}, P_{t+1}, A_{t+1})$. However, due to inherent biological complexity and variation we cannot expect this to be true. The LPA model does not take into account all biological factors related to the beetle. Processes such as birth rates and death rates may vary across a population and fluctuate with time, and so we will not be able to find a set of parameter values under which the model exactly matches the data. The best we can hope for is parameter values which give us model predictions closest to the observed data.

We need to build this variation, or "noise," into our model in order to account for the variability in the data. In order to do this, we need to know more details about the noise, also known as stochasticity. The two main types of noise are environmental and demographic. The main source of demographic noise is variation in model parameters,

such as birth and death rates, within a population. Environmental stochasticity reflects disturbances affecting the entire population, such as fluctuations in weather. Demographic noise typically predominates at low population levels, while environmental noise is often the main source of variation at high population levels.

3.3 Maximum Likelihood Estimation

The statistical method of maximum likelihood estimation was used to estimate parameter values for the LPA model. We will illustrate the procedure for a simpler model [1]. The general form of a one dimensional discrete time map is given by

$$x(t+1) = f(x(t), \theta_1, ..., \theta_p) \tag{2}$$

where $\theta_1, ..., \theta_p$ are p model parameters. Two examples commonly used to model populations are the Ricker map

$$x(t+1) = bx(t)\exp(-cx(t)) \tag{3}$$

and the discrete logistic map

$$x(t+1) = \frac{bx(t)}{1+cx(t)} \tag{4}$$

where b is the birthrate and the parameter c measures the intensity of density effects.

With all parameter values assigned, model maps (2) $x(t)$ to $x(t+1)$. In order to estimate model parameter values from data, we must make assumptions about the type of noise present in the system, since different types of noise are built into the model in different ways. Noise is added as a normal random variable with mean 0 and variance v, and it can be shown (see part II) that for demographic noise this variable should be added on the square root scale while for environmental stochasticity it should be added on a logarithmic scale (in order to best keep the variance constant). The following is a stochastic version of equation (2) with environmental noise:

$$x(t+1) = f(x(t), \theta_1, ..., \theta_p)\exp(E_t) \tag{5}$$

The environmental stochastic Ricker model is given by

$$x(t+1) = bx(t)\exp(-cx(t))\exp(E_t) \tag{6}$$

and the environmental stochastic discrete logistic model is given by

$$x(t+1) = \frac{bx(t)}{1+cx(t)}\exp(E_t). \tag{7}$$

Models (5), (6), and (7) are examples of nonlinear, autoregressive models. Assume we have a data set of q points $(y_1, y_2, ..., y_q)$ and we want to parameterize model (5). We first want to transform both sides of equation to a log scale (5), so noise is additive:

$$\ln x(t+1) = \ln f(x(t), \theta_1, ..., \theta_p) + E_t. \tag{8}$$

The goal is to find the set of parameter values $\theta_1, ..., \theta_p$ that maximize the probability, or likelihood, that the observed data set $(y_1, y_2, ..., y_q)$ would be generated from this stochastic model.

This probability is given by the likelihood function L (which we want to maximize):

$$L(\theta_1, ..., \theta_p, v) = \prod_{t=1}^{q} p(w_t|w_{t-1}) \tag{9}$$

where $w_t = \ln y_t$, and $p(w_t|w_{t-1})$ is the probability that w_t occurs given w_{t-1} occurs. This probability has a normal distribution with mean $\ln f(y_{t-1}, \theta_1, ..., \theta_p)$ and variance v:

$$p(w_t|w_{t-1}) = \frac{1}{\sqrt{2\pi v}}\exp(-\frac{1}{2v}(w_t - \ln f(y_{t-1}, \theta_1, ..., \theta_p))^2) \tag{10}$$

Then

$$L(\theta_1, ..., \theta_p, v) = \prod_{t=1}^{q} \frac{1}{\sqrt{2\pi v}} \exp(-\frac{1}{2v}(w_t - \ln f(y_{t-1}, \theta_1, ..., \theta_p))^2). \tag{11}$$

Maximum likelihood parameter estimates are those values of $\theta_1, ..., \theta_p, v$ that maximize equation 11. It is equivalent, and often easier, to maximize $\ln L(\theta_1, ..., \theta_p, v)$ which we will denote by $l(\theta_1, ..., \theta_p, v)$.

$$l(\theta_1, ..., \theta_p, v) = -\frac{q}{2}\ln(2\pi) - \frac{q}{2}\ln v - \frac{1}{2v}\sum_{t=1}^{q} r_t^2(\theta_1, ..., \theta_p) \tag{12}$$

where

$$r_t(\theta_1, ..., \theta_p) \doteq \ln y_t - \ln f(y_{t-1}, \theta_1, ..., \theta_p) = \ln(\frac{y_t}{f(y_{t-1}, \theta_1, ..., \theta_p)}) \tag{13}$$

and are known as the "log-residuals."

The critical points of l can be found by solving the following equations for $(\theta_1, ..., \theta_p, v)$:

$$\sum_{t=1}^{q} r_t(\theta_1, ..., \theta_p) \frac{\partial_{\theta_i} f(y_{t-1}, \theta_1, ..., \theta_p)}{f(y_{t-1}, \theta_1, ..., \theta_p)} = 0 \tag{14}$$

$$v = \frac{1}{q}\sum_{t=1}^{q} r_t^2(\theta_1, ..., \theta_p) \tag{15}$$

In general, these equations can be very complicated and must be solved numerically with a computer.

3.4 Assessing the fit of a model

Once we have parameter estimates, we can ask how well a model accounts for data, such as population census data. We also may have two competing models and may want to use the one that best describes our data set. How do we choose? Figure 1 shows a time series of 101 data points ($q = 100$). The maximum likelihood parameter estimation formulas gives estimates of

$$b \approx 7.591, \quad c \approx 0.009751, \quad v \approx 0.009702 \tag{16}$$

for the environmental stochastic Ricker model. This data can also be used to calculate parameter values for the environmental stochastic discrete logistic model. Maximum likelihood estimates are

$$b \approx 13.14, \quad c \approx 0.06059, \quad v \approx 0.08163. \tag{17}$$

The difference between the one-step prediction of the parameterized model and the actual next data point is known as a residual. Since we added noise in the form of a Normal random variable with mean 0 and variance v to our model on a log scale, the log-residuals should be normally distributed with mean 0 and variance v. Figure 2 shows a histogram of the log-residuals for the stochastic logistic model (a) and the stochastic Ricker model (b). Both histograms have a mean near 0. The stochastic Ricker log-residuals appear to be closer to a Normal distribution. The models also differ in their "R-squared" value, which tells us the fraction of variability in the data that is explained by the model.

$$R^2 \doteq 1 - \frac{v}{v_0} \tag{18}$$

where v is given by equation (15) and v_0 is the variance of the transformed data $w_t = \ln y_t$. The environmental stochastic Ricker model explains 94.89% of the variability in the data set 1, while the environmental stochastic discrete logistic model explains only 57.02%.

Once we have seen that the Ricker model can account for the data used in its parameterization, an additional test is to look at the distribution of the log-residuals for additional data left out of its parameterization. If the log-residuals are normally distributed with the correct mean and variance, then this serves as evidence of the predictive power of

Figure 1:

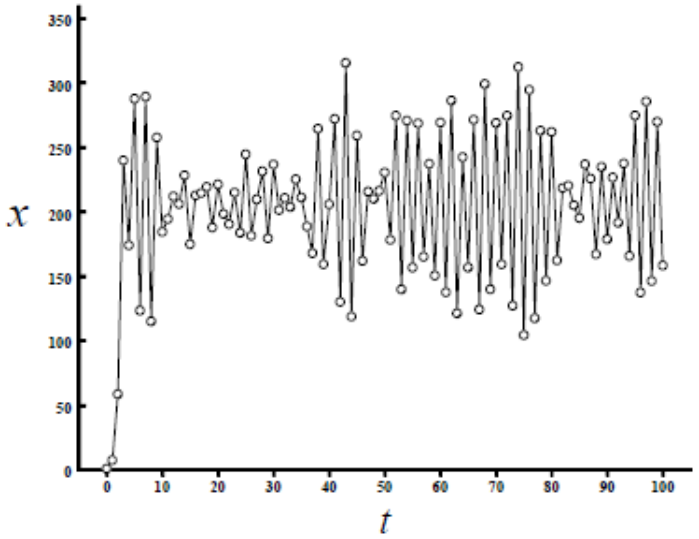


Figure 2:

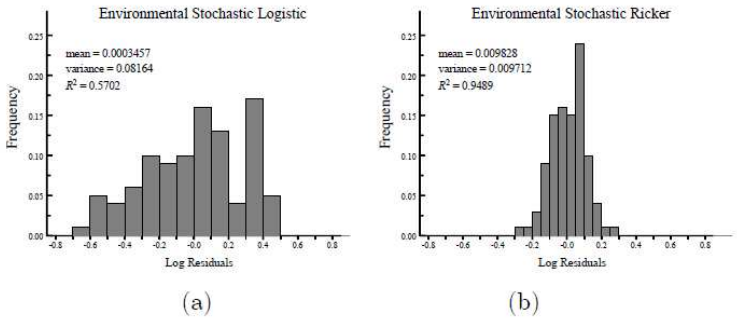
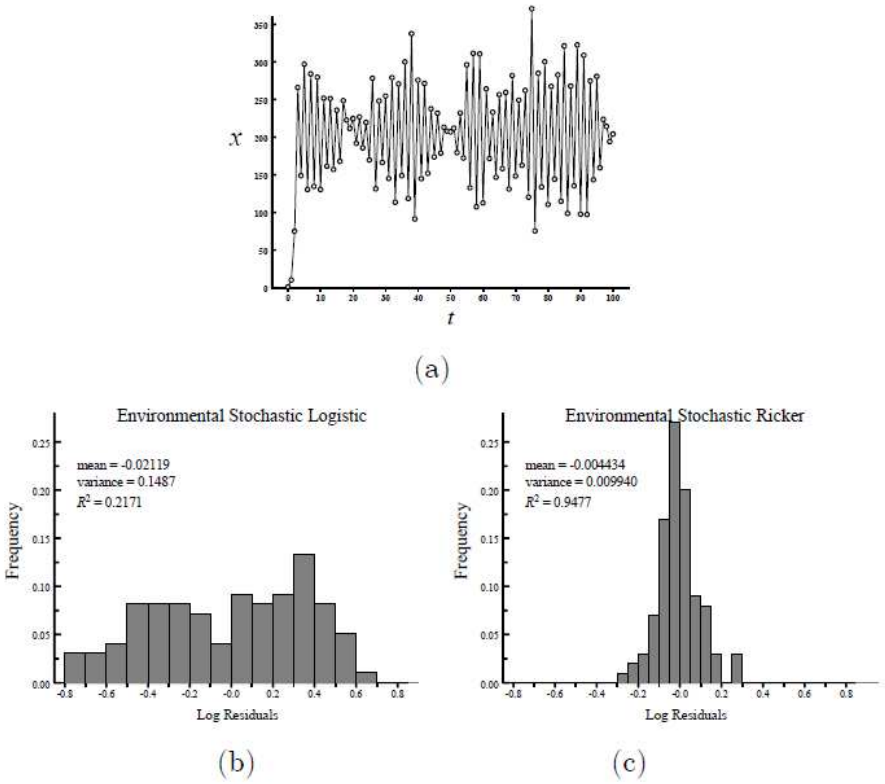


Figure 3:

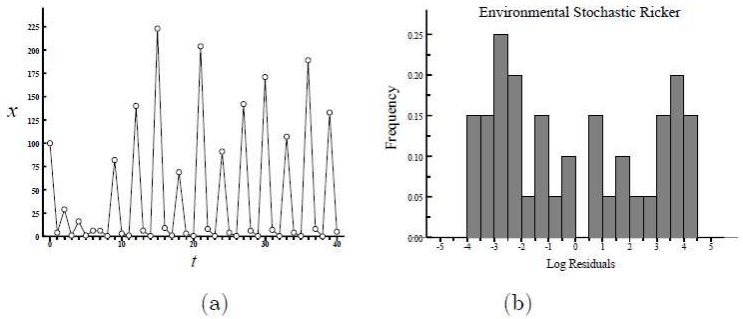

x
t
(a)
Environmental Stochastic Logistic
mean = -0.02119
variance = 0.1487
R^2 = 0.2171
Frequency
Log Residuals
(b)
Environmental Stochastic Ricker
mean = -0.004434
variance = 0.009940
R^2 = 0.9477
Frequency
Log Residuals
(c)

the model. Data from another replicate of the experiment that produced figure 1 is shown in figure 3, (a). This data is much more oscillatory than that from the first experiment. Residuals are calculated for the stochastic discrete logistic and stochastic Ricker models with parameter estimates (**??**) are . The histograms of log-residuals are shown in 3. We see the residuals for the stochastic discrete logistic model are much less convincing of a normal distribution than the stochastic Ricker model. Moreover, the R^2 value of the stochastic discrete logistic model dropped to 21.71% while the R^2 value of the stochastic Ricker model was essentially unchanged at 94.77%. We reject the fit of the stochastic discrete logistic model to this data. By doing this, we are rejecting either the deterministic discrete logistic model or the type of noise structure added to it.

3.5 Modeling Tribolium

Figure 4(a) shows a data set of adult *Tribolium* census counts. The maximum likelihood parameterization process described above was used to find maximum likelihood parameter estimates for the environmental stochastic Ricker model. A histogram of the one-step log residuals is plotted next the the data in 4(b). These histograms are far from normal and we reject this model for this data. For the case of *Tribolium*, the one-dimensional Ricker model cannot

Figure 4:



take into account the known nonlinear interactions between life-stages. While it may be adequate for some *Tribolium* data sets, we see it is insufficient for others such as 4(a). The maximum likelihood techniques described above can be extended to multivariate models such as the LPA model, which is better suited to describe *Tribolium* dynamics. Details are given in [3, 4].

4 Stochastic Models - Mathematical Details

4.1 Nonlinear Autoregressive Models

NLAR models are a class of stochastic models that add a random term to the deterministic (nonrandom) model skeleton on a scale at which the variance of the random term is approximately constant. For example, consider a deterministic 1-d model

$$x_{t+1} = f(x_t)$$

with a time series of data points y_i. A transformation of data points is given by $w = g(y)$ and of model predictions $n = g(x)$. Then our stochastic model takes the form $g(x_{t+1}) = f(g(x_t)) + E_t$ or $n_{t+1} = f(n_t) + E_t$, with E_t having mean 0 and constant variance v (independent of t). E_1, E_2, ... are independent random variables.

The transformation which stabilizes the variance of E_t depends on the type of noise we are assuming is dominating the system. Consider a simple survival process where μ is the death rate. The deterministic model is given by:

$$x_{t+1} = (1 - \mu)x_t$$

Now suppose we want to add environmental stochasticity. Then the death rate itself, μ, is a random variable (as is $1-\mu$). Thus x_{t+1} is a random variable that depends on x_t. We have

$$Mean[x_{t+1}] = Mean[1-\mu]x_t$$

and

$$Var[x_{t+1}] = Var[1-\mu]x_t^2 = Var[\mu]x_t^2.$$

If we want to include demographic stochasticity, then each individual present at time t (x_t independent individuals) survives to time $t+1$ with probability $1-\mu$. Then the number of individuals surviving to time $t+1$ is a binomial random variable with x_t trials and probability of success $1-\mu$. We have

$$Mean[x_{t+1}] = (1-\mu)x_t$$

and

$$Var[x_{t+1}] = (1-\mu)\mu x_t.$$

Now we want to find the transformation to use in our stochastic model. As stated above, $n_t = g(x_t)$. We want $\frac{dg(x)}{dx} > 0$. The first order Taylor approximation of n_{t+1} is given by

$$n_{t+1} = g(x_{t+1}) \approx g(x_t) + g'(x_t)(x_{t+1} - x_t) \tag{19}$$

Here we consider x_{t+1} a random variable conditioned on a given value of x_t. Taking the variance of both sides of equation 19

$$\begin{aligned} Var[n_{t+1}] \approx Var[g(x_t) + g'(x_t)(x_{t+1} - x_t)] \\ &= Var[g'(x_t)(x_{t+1} - x_t)] \\ &= (g'(x_t))^2 Var[(x_{t+1} - x_t)] \\ &= (g'(x_t))^2 Var[x_{t+1}] \end{aligned} \tag{20}$$

Assuming the variance of x_{t+1} is a function of x_t, let

$$Var[x_{t+1}] = v(x_t)$$

so that

$$Var[n_{t+1}] \approx (g'(x_t))^2 v(x_t)$$

We need to find the transformation $g(x)$ that makes $Var[n_{t+1}]$ approximately constant. This amounts to solving the following

$$(g'(x))^2 v(x) = c_0$$

for $g(x)$ with constant c_0. The solution is given by

$$g(x) = \int (\frac{c_0}{v(x)})^{\frac{1}{2}} dx + c_1$$

with arbitrary constant c_1.

We can substitute back in for $v(x)$ the variances we found for x_{t+1} under environmental and demographic stochasticity to find the appropriate $g(x)$. With $c_0 = Var[\mu]$ and $c_1 = 0$, we see that a transformation of $g(x) = \ln x$ stabilizes variance in the environmental case, while $c_0 = (1-\mu)\mu/4$ and $c_1 = 0$ leads to a stabilizing transformation of $g(x) = \sqrt{(}x)$ in the demographic case.

5 Maximum Likelihood Estimation - Calculating Parameter Values

The maximum likelihood estimation procedure was described in Part I. In general, likelihood functions can be very complicated and must be maximized numerically with a computer. An example of a model for which we can calculate explicitly the maximum likelihood parameter estimates is the Ricker Map with environmental stochasticity from Part I, where $\theta_1 = b$ and $\theta_2 = c$. Parameter estimates are calculated to be:

$$\begin{aligned}
\ln b &= \frac{s_1 s_4 - s_3 s_2}{s_1^2 - q s_2} \\
c &= \frac{q s_4 - s_1 s_3}{s_1^2 - q s_2} \\
v &= \frac{1}{q}(q \ln^2 b - 2c \ln b s_1 + c^2 s_2 - 2 \ln b s_3 + 2c s_4 + s_5)
\end{aligned}$$

where

$$\begin{aligned}
s_1 &\doteq \sum_{t=1}^{q} y_{t-1} \\
s_2 &\doteq \sum_{t=1}^{q} y_{t-1}^2 \\
s_3 &\doteq \sum_{t=1}^{q} \ln(\frac{y_t}{y_{t-1}}) \\
s_4 &\doteq \sum_{t=1}^{q} y_{t-1} \ln(\frac{y_t}{y_{t-1}}) \\
s_5 &\doteq \sum_{t=1}^{q} \ln^2(\frac{y_t}{y_{t-1}})
\end{aligned}$$

These formulas can be tested by setting $v = 0$ and iterating the Ricker map with set parameter values to generate a data set y_t. Then using y_t in the above equations, we see that we recover our set parameter values. This should make sense, as the parameter values most likely to generate the data set y_t are the ones that we actually used to generate y_t.

All of the data sets, figures, and computations for this project were taken from [1] and [2].

References

[1] CUSHING, J. M. *An Introduction to Structured Population Dynamics*, vol. 71. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1998.

[2] CUSHING, J. M. Matrix models and population dynamics. In *Mathematical Biology, IAS/Park City Mathematics Series* (Providence, RI, 2009), J. Keener and M. Lewis, Eds., American Mathematical Society.

[3] DENNIS, B., DESHARNAIS, R. A., CUSHING, J. M., AND COSTANTINO, R. F. Nonlinear demographic dynamics: Mathematical models, statistical methods, and biological experiments. *Ecological Monographs 65* (1997), 261–281.

[4] DENNIS, B., DESHARNAIS, R. A., CUSHING, J. M., HENSON, S. M., AND COSTANTINO, R. F. Estimating chaos and complex dynamics in an insect population. *Ecological Monographs 71* (2001), 277–303.