www.simiode.org
SIMIODE Systemic Initiative for Modeling Investigations and Opportunities with Differential Equations

STUDENT VERSION
DISEASE SPREAD

Brian Winkel
Director SIMIODE
Cornwall NY USA

STATEMENT

We offer a modeling opportunity in which the phenomenon of the spread of disease can be described by one differential equation.

First, let us consider growth in which we have a quantity which grows by a constant amount, say, k, i.e.,

$$y'(t) = k \ , \ y(0) = y_0 \tag{1}$$

In this case we can integrate both sides to obtain $y(t) = k \cdot t + c$ where the initial condition $y(0) = y_0$ determines c so that we have a complete solution, $y(t) = k \cdot t + y(0)$.

A more interesting growth equation comes from assuming that the growth rate of some quantity is proportional to the amount present, e.g., money earns interest and populations breed. Here

$$y'(t) = k\, y(t) \ , \ y(0) = y_0 \tag{2}$$

Exponential Growth Population Modeling

Exponential growth (and its related decay issues) stem from the following differential equation:

$$y'(t) = ky(t) \ , \ y(0) = y_0 \ . \tag{3}$$

k is referred to the *rate constant* and if $k > 0$ we call it a growth constant, while if $k < 0$ we call it a decay constant.

This exponential growth model can be used to predict the growth of money where k is the decimal value of the percent interest on continuously compounded interest situations or the growth of organisms where k is the per unit time growth rate. In terms of parameter estimation the

problem of estimating k is universally the same and the various approaches (1) linearizing the solution by taking natural logarithms of the non-time variable and fitting the linearized data using linear regression, (2) estimating growth rates and fitting a linear model relating $y'(t)$ and $y(t)$, or (3) direct nonlinear fitting of the analytic solution to (3) to the data, will serve well to estimate k. We seek more appropriate models for growth in biological situations and this demands we do a better job of modeling the reality of finite resources or populations. This leads to a limited growth population model.

Limited Growth Population Model

Time (days)	Population (#)	Time (days)	Population (#)
0	200	26	783
2	240	28	792
4	281	30	837
6	317	32	800
8	362	34	866
10	392	36	933
12	437	38	942
14	496	40	941
16	542	42	986
18	601	44	964
20	615	46	940
22	662	48	1009
24	758	50	984

Table 1. Population data with units of number (#) of members of the population at time t in days.

In Table 1 we have data from a population which does not actually grow exponentially. There is a limit to growth. Often this is a natural limit, such as regional resource limitations, geometry of the environment, or borders. Just consider growing microorganisms in a Petrie dish which has only a finite space.

Conjecturing a growth model for Table 1 data

1) Take time to consider how we could modify the differential equation (3) for exponential growth so as to actually model growth that would level off or whose growth rate would eventually get smaller and smaller. Discuss your models with colleagues and decide if any of the models under discussion are suitable. Look for characteristics of the differential equation you offer. For

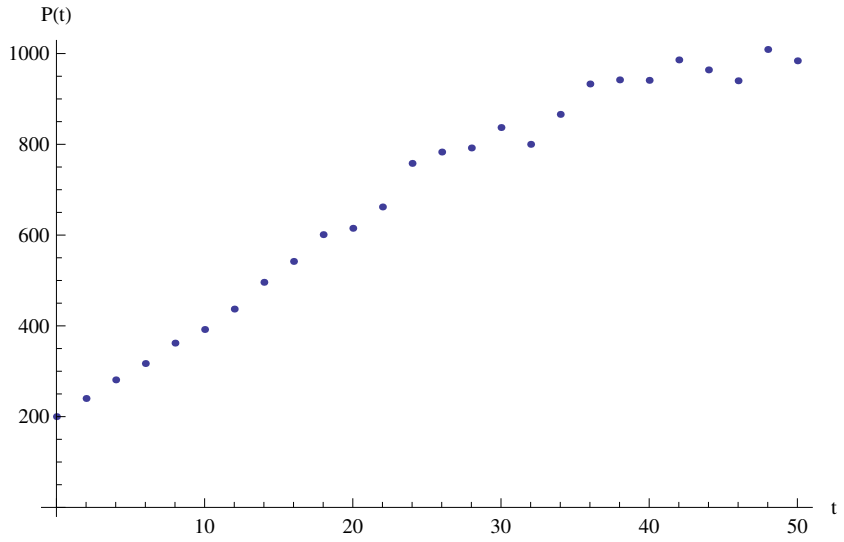

Figure 1. The plot of the population data in Table 1.

example, does your differential equation appear to come to a point where growth is 0 (leveling off)? Does it have any flaws? For example, does it create some population out of nothing?

2) Produce some conjectures and "talk" about other models being proposed by your colleagues. Probe each model, test it under obvious conditions, e.g., "If we have no population do we get no growth in our model?"

While not ignoring your efforts we are going to turn - as you turn to the next page - to a well-established model for a while and we will offer up a chance for you to try out your model shortly. So hold on to your ideas. See the last question in this paper in which you can take your model as far as you wish.

The differential equation (4) is referred to as the *logistic differential equation*. It is used in such areas as modeling limited growth population biology as well as the study of the spread of rumors, disease, and technologies:

$$y'(t) = ry(t)\frac{K - y(t)}{K}, \qquad y(0) = y_0\,. \tag{4}$$

Here $y(t)$ is the population in appropriate units.

Scrutinize the logistic differential equation model

3) Take some time to give the logistic model the scrutiny you and your colleagues gave to each other's models. Try to push it around a bit, play some "what if" games with it. Test it, interpret it, challenge it. Well, what do you think about this model?

In different contexts the logistic differential equation (4) could be the percent of the population that has heard a specific rumor or acquired knowledge or uses a specific technology. Or it could be the number in the population that have had a given disease. r represents the intrinsic growth rate (in population models r is birth rate less death rate) with K as the limiting value of $y(t)$. K is called the *carrying capacity*. The latter comes from ecology and refers to the number of the given species that the environment can support or carry. For excellent histories of the logistic curve in the development of population biology see [2, 3].

Now, let us return to the data as offered in Table 1. This particular data is what we call a *toy* set of data and is not from any real experiment. In later work with the logistic model we shall use actual data from [1]. Indeed, in this activity you will generate data from a physical simulation using M&M candies as members of your population.

Upon plotting the data (see Figure 1) and developing a model, presumably the logistic differential equation (4), based on the sigmoidal or "S" shape of the solution of the logistic equation (4), we ask how we might confirm or deny that (4) is a good model for our data. Throughout this early discussion we may never actually solve (4) by hand - although one could do so by using the separation of variables and partial fraction strategies described elsewhere. Rather, when we do need an analytic solution we rely on *Mathematica*'s symbolic solving command `DSolve` or we may go to its numerical solving command `NDSolve` as appropriate.

In assessing how good a model we may have developed in (4) we might consider determining criteria for estimating the parameters, i.e., finding the parameters which when we substitute them in (4) and solve it will permit us to compare how our model predicts the data in Table 1.

4) Given the data in Table 1 and as presented in Figure 1 how would you confirm or deny that (4) is a good model for our data, for the phenomena of limited growth.
5) Before going on, consider how you would estimate the parameters r and K and offer up your ideas and approaches. Discuss the pros and cons of your approaches and your colleagues' approaches. Write them down. State the steps you might use. Do it here, do it now.

Developing strategies for estimating parameters

Here are our several approaches that usually crop up. How do your approaches compare?

1. Can we estimate the parameters by transforming the logistic differential equation model and then determining the least sum of square differences between transformed data and model to determine which parameters give the best fit - nonlinear fit of $\frac{dP}{dt}$ vs. $rP\frac{K-P}{K} = aP - bP^2$, where $a = r$ and $b = \frac{r}{K}$?
2. Can we estimate the parameters by transforming the ODE model and then determining the least square differences between transformed data and model to determine which parameters give the best fit - linear fit of $\frac{1}{P}\frac{dP}{dt}$ vs. $r\frac{K-P}{K} = a - bP$, where $a = r$ and $b = \frac{r}{K}$?
3. Can we estimate the parameters by analytically solving the logistic differential equation model in (4) and then determining the least sum of square differences between data and analytic model to determine which parameters which yield the least sum of square differences, i.e. nonlinear fit of the analytic solution to the data at the time observation points of the data?
4. Can we estimate the parameters by numerically solving the logistic differential equation model for a wide variety of parameters and determining the least sum of square differences between data and model values for EACH set of parameters and then determine the parameters r and K which give rise to the smallest least sum of square differences?

Approach (4) is a bit strange and rather computationally intense, but it is still doable, namely try lots, and we mean may LOTS, of parameter sets r and K in a reasonable vicinity of what you believe might work to see if you can get systematically lucky and get pretty good estimate.

However, let us concentrate on the first two approaches (1) and (2). Even inside approaches (1) and (2) we have several approaches to transforming the data. For example in each case we can take a first difference approach to estimating the derivative from the data or a symmetric first difference approach to estimating the derivative. Thus, we shall have approaches (1a) and (1b) as well as (2a) and (2b) with the (a) denoting the plain first difference approach and (b) denoting the symmetric first difference approach.

In all approaches we shall use the same measure of how good the method is, namely the sum of square errors, i.e., the squares of the difference between the value of the solution to the logistic differential equation (4) evaluated at the time value for each data point and the corresponding population value at that time. This would look like:

$$\sum_{i=1}^{n} (\hat{y}(t_i, r, K) - P_i)^2 \,,$$

where $\hat{y}(t_i, r, K)$ is the value of our solution to our logistics differential equation (4) with parameters (r, K) at time t_i and P_i is the value of the population at time t_i from our data set for $i = 1, \ldots, n$, while n is the size of our data set.

Parameter estimation opportunities

6) Use approach (1a) and (1b) to estimate parameters r and K using Trendline in EXCEL and fitting a parabola with intercept zero to the transformed data. Use your parameters in each case and return to *Mathematica*. Enter the data as a list of pairs of numbers and `ListPlot` that data, calling the plot, dataPlot. Now solve the differential equation and grab the solution with this line of *Mathematica* code:

```
ys[t_] = y[t] /.DSolve[y'[t] == r y[t] (K - y[t])/K, y[0] == y0,
        y[t], t][[1]]
```

with the parameters from your EXCEL adventure. Your solution should look like this, only with your numerical estimates for r and K from your EXCEL adventure, instead of a general from equation with the letters r and K as in (5):

$$ys(t) = \frac{Ky_0e^{rt}}{K + y_0e^{rt} - y_0}\,. \tag{5}$$

Now after substituting $y_0 = 200$ (see Table 1) into (5) we proceed to evaluate the solution at each time $t = 0, 2, 4, \ldots, 50$ from the data; subtract the observed corresponding data values, 200, 240, 281, ..., 984; square these differences; and add the square differences or errors to obtain:

$$ss(r, K) = \sum_{i=1}^{n} (\hat{y}(t_i, r, K) - P_i)^2\,. \tag{6}$$

We seek the values of r and K that will minimize the sum in (6) and so we call up *Mathematica*'s `FindMinimum` command. From this output we can read our best r and K values. Take these and substitute them into (5) along with $y_0 = 200$ and we have a fully developed model which we can plot (call it ModelPlot) and compare with our data plot by `Show`ing them together in *Mathematica*.

Discuss your results from (1a) and (1b) approaches.

7) Use approaches (2a) and (2b) to estimate parameters r and K using Trendline in EXCEL and fitting a linear function to the transformed data. Use your parameters in each case and return to *Mathematica*. Enter the data as a list of pairs of numbers and `ListPlot` that data, calling the plot, dataPlot. Then solve the differential equation and grab the solution with this line of *Mathematica* code:

```
ys[t_] = y[t] /.DSolve[y'[t] == r y[t] (K - y[t])/K, y[0] == y0,
        y[t], t][[1]]
```

with the parameters from your EXCEL adventure. Your solution should look like this:

$$ys(t) = \frac{Ky_0e^{rt}}{K + y_0e^{rt} - y_0}\,. \tag{7}$$

After substituting $y_0 = 200$ (see Table 1) into (7) we proceed to evaluate the solution at each time $t = 0, 2, 4, \ldots, 50$ from the data; subtract the observed corresponding data values,

200, 240, 281, ..., 984; square these differences; and add the square differences or errors to obtain:

$$ss(r, K) = \sum_{i=1}^{n} (\hat{y}(t_i, r, K) - P_i)^2 \,. \tag{8}$$

We seek the values of r and K that will minimize the sum in (8) and so we call up *Mathematica*'s `FindMinimum` command. From this output we can read our best r and K values. Take these and substitute them into (7) along with $y_0 = 200$ and we have a fully developed model which we can plot (call it ModelPlot) and compare to our data plot by `Show`ing them together in *Mathematica*.

Discuss your results from (2a) and (2b) approaches.

8) Finally, directly estimate r and K by first evaluating the solution, (7), to the logistic differential equation (4) at each time value $t = 0, 2, 4, \ldots, 50$ along with $y_0 = 200$ and form the sum of square errors, as before. Except that now we can simply `FindMinimum` this sum of square errors to find directly which r and K make the solution ((7)) with these r and K values our best fitting model.

Take these values of r and K and substitute them into (7) along with $y_0 = 200$ and we have a fully developed model which we can plot (call it ModelPlot) and compare to our data plot by `Show`ing them together in *Mathematica*.

Discuss your result and compare it with those from (1a) and (1b) as well as from (2a) and (2b) approaches.

Modeling spread of disease with M&M simulations

While we will examine some real data from the literature and attempt to model it with a logistic model, we take some time here to develop our own data of the spread of disease using this logistic model approach. Namely, if in a population of size K we have $y(t)$ members who have a communicable disease and can spread it upon contact with susceptible members of which there are $(K - y(t))$ then a possible model for the rate of the spread of the disease could be the logistic equation:

$$y'(t) = \alpha y(t)(K - y(t)) \tag{9}$$

where the product $y(t) \cdot (K - y(t))$ represents the number of interactions that could take place per unit time and α is a constant which reflects just how many interactions take place AND how many result in the spread of the disease. Of course we could convert to percentage of the population whereupon $K = 1$ stands for 100% of the population, $y(t)$ stands for the percentage of the population who have the disease, and $K - y(t) = 1 - y(t)$ stands for the population which does not have the disease and is susceptible.

We recognize (9) as an over simplified model of the spread of disease, because it does not reflect several realities, among them the need to accommodate immune individuals, the various stages

of gestation of the disease before it might be transmittable, effects on various age groups, etc. Nevertheless, we shall proceed to work with the model of (9), picking up on more realistic variations as our modeling and differential equations skills increase!

9) Articulate the assumptions and oversimplifcations we need to make to postulate (9) as a model for the spread of disease.

So let us simulate a disease spread among M&M's!!! We shall need a regular size bag of M&M chocolates - not the peanut kind (there should be about 55 pieces in this size bag); a small cup or glass to hold the M&M's; and a grid (copy in Figure 2 blown up to a size (300%) so that each grid is 1.32 cm high and wide - the radius of an M&M is 1.32 cm. Indeed, in the folder for this scenario there is a pdf file, M&MLayout.pdf which, if printed, is the exact size of your population surface which you will need. Otherwise, follow instructions in caption of Figure 2.

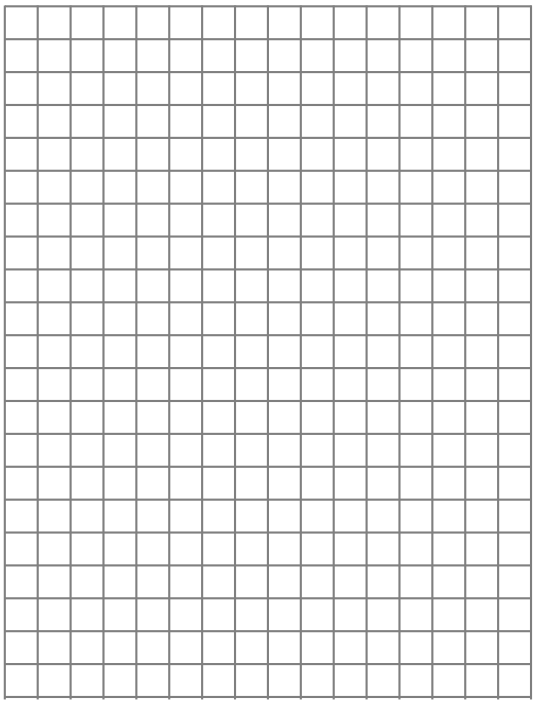

Figure 2. A grid used for our simulation is to be reproduced so that each grid is 1.32 cm high and wide when expanded to 300% - the radius of an M&M is 1.32 cm. To make the necessary grid sheet for the simulation copy this actual grid on a copy machine at 300% - making sure it exactly fills an 8.5" x 11" sheet.

The Simulation Experiment

We are going to simulate the spread of a disease. Prepare and then cut out a grid of squares each 1.32 cm by 1.32 cm where 1.32 cm is the diameter of an M&M on an 8.5" x 11" sheet of paper. See

instructions in Figure 2. Each grid box represents an individual M&M. We put bold X's in 8 of the boxes to indicate that the disease initially has 8 individual M&M's infected. Select these randomly. If you are not confident in doing so then use these *Mathematica* instructions to select randomly

```
Table[{Random[Integer, {1, 16}], Random[Integer, {1, 21}]}, {i, 1, 8}]
```

where there will be 16 M&M's width and 21 M&M's height.

Now tape one edge of the 8.5" x 11" sheet of paper to a desk or table flat surface and then line the sheet with four vertical walls on each edge, e.g., four reams of 8.5" x 11" paper to make a bounded fenced region.

Take the 55 M&M's in a standard M&M bag and place them in a cup. These form our healthy population, the susceptibles. Thus we have 63 M&M's in our population - 8 infected and 55 susceptibles - at the start of our simulation.

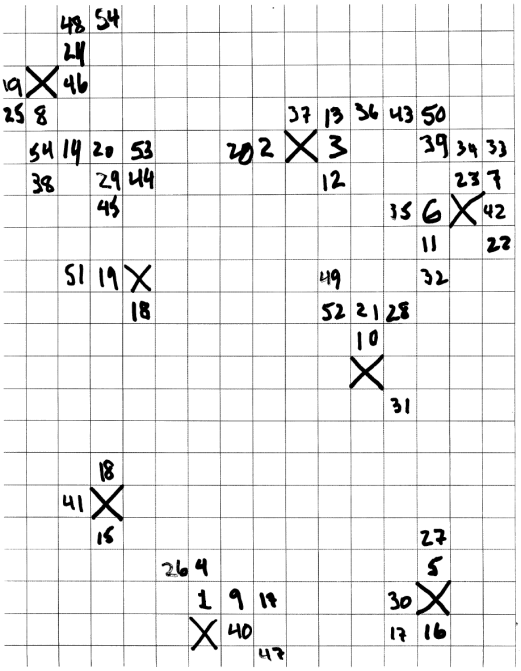

Figure 3. A grid used for our first simulation of infecting M&M's. Notice the 8 original infectives with an X in each box and then the consecutive infectives numbered 1-55.

Gently toss the M&M's from the cup "randomly" on the grid in the bounded fenced region. Each time we toss the M&M's we will consider a generation or cycle or unit of time. If an M&M "touches" (overlaps in any part) a box with an X in it we mark the box in which most of the M&M resided with a number (starting with 1 for our first infected M&M beyond the original 8, and then a 2 for our second, and so forth). As an M&M becomes infected we remove it from the healthy population, noting that our number of infecteds is in addition to the original or current infected population. The newly numbered boxes stand for infected M&M's who stand ready (and willing!)

to infect other susceptible M&M's should one of these susceptible M&M's touch one of the infected M&M squares. See Figure 3 for a summary simulation.

Figure 4 (in Black and White) shows the result of the simulation at some point in time along the process.

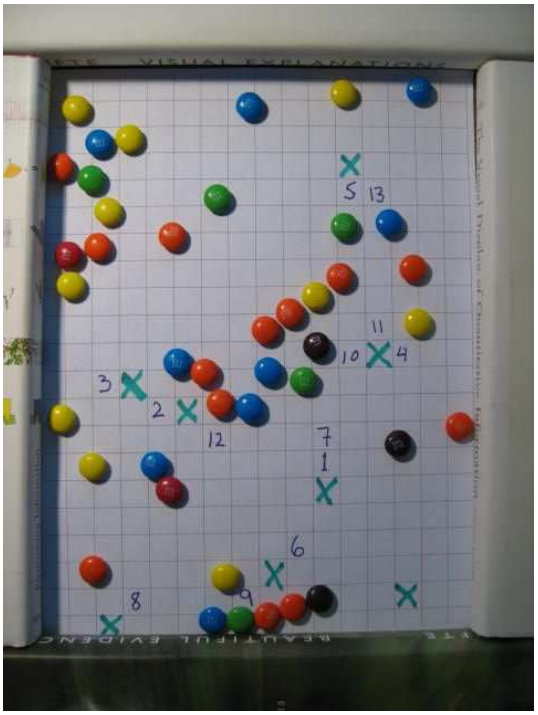

Figure 4. Photograph of actual grid and M&M's at some point in time of the simulation. Can you identify the newly infected M&M's from this picture?

At each generation we toss the cup, determine those M&M's who get infected, increment the number of newly infecteds, and write that number in a square nearest to the M&M that just got infected. We also keep a tally sheet of the total number of infected at each new generation or cycle or time interval. We show such numerical results in Table 2.

The "touching" or overlap simulate contact and we presume contact is sufficient to transmit the disease, i.e. absolute contagion. Of course, we could toss a die, for example, to determine for each touch what the chance of spread of the disease could be, using some preassigned rule such as if face of die shows 5 or 6 then there is no infection in this case; doing this for each touched situation. This would slow up the process, but make it more realistic in some sense as most diseases are not automatically transmitted upon touch.

10) Either from the data in Table 2 or from one of your own experiments analyze the data by assuming the logistic differential equation (10) could be used to describe the spread of the

Generation	# Infected
0	8
1	11
2	16
3	22
4	28
5	37
6	46
7	48
8	54
9	58
10	61
11	62
12	62
13	62
14	62
15	63

Table 2. Data from our first experiment in simulating infection of M&M's. Yours should look somewhat the same, don't you think?

disease between the susceptible ($y(t)$) and infected ($K - y(t)$) M&M's.

$$y'(t) = ry(t)\frac{K - y(t)}{K}, \qquad y(0) = y_0\,. \tag{10}$$

Then use some of our previously discussed methods to determine the parameters r and K.

11) Discuss the fact that if you do not assume $K = 63$, i.e. there are eventually at most $8 + 55 = 63$ M&M's in our population (the original 8 infectives and the 55 susceptibles from your bag of yummy M&M's) you might get a different model than if you do NOT assume a value for K.

12) Discuss (and do!) variations of the model, such as

- making the squares on your grid larger or smaller;
- increasing the number of original infectives;
- do not put the new infectives where they were infected but rather randomly put them in some other position (they could have wandered a while before dying);
- making a grand sheet of taping four 8.5" x 11" sheets to form a large 17" x 22" sheet with the same number of initial infecteds and the same number of susceptibles; or
- making your own changes.

13) What is the value of r that you get in each case? Does it change in any way? What does r mean? What is the significance of r without the limiting value K (i.e. unlimited bag of M&M's - yummy, yummy!!!)

14) A time intensive (possibly project) activity is to use some random feature to determine that IF an uninfected or susceptible M&M comes in contact with an infected one then with some probability the disease is passed on – not all contact results in the spread of the disease in this case. For example, use a six sided die and consider each contact, one by one. Toss the die, and if, say, it is a 3 or a 5 (you make up your own rule), then an infection actually takes place; otherwise no infection takes place and the potentially infected M&M goes back into the susceptible pool. Change your probability rule. Does increasing the chance of infection at each step do anything to the data, to the plot, to r?

15) Take your model or one of those you develped before we "spilled the beans" (M&M's actually!) and shared the logistic differential equation and address questions (3) - (7) above with that model, addressing the issues as best you can.

COMMENTS

We have never done all of these activities in one "setting." Rather, we have used bits and pieces, broad swatches of this material in various situations and students are very good at addressing how we might estimate parameters in the logistic differential equation, probably because they are familiar with EXCEL and differencing and applying functions (e.g., logarithm) to whole columns of numbers is quite familiar to them.

When it comes to building models this is an early activity to try to build in limited growth and the students really struggle with this. Often, I give them lots of time, say 30 minutes with some individual and small group interaction with me, as well as total class time in which we examine some candidates and teach how to test at some obvious points, e.g., when the population is 0 does the model in question say there is growth - not good! Most times, indeed, almost always the class will come around to some form of the logistic differential equation and feel pretty confident about the group effort. This is good, for when we actually turn the page and offer them the same model and tell them that historically many mathematicians and ecologists have come to the same conclusion they are pleased with themselves.

This is an early activity in which we go after parameters through estimation techniques and so we take time to compare our results and methods. We depend heavily on technology, Mathematica and EXCEL and we are often teaching students the syntax en route to modeling success, so patience is the order of the day. However, since we do lots of parameter estimation efforts in our course this pays off often.

Instead of offering comments here we address each of the Activities offered in a separate Mathematica notebook, both in .nb and .pdf format and offer detailed comments on mathematics, technology use, and student efforts.

NB: With regard to the set of references below, what are the chances that all reference lead authors would have a last name beginning with letters only from the first half of the alphabet? Just a curiosity we noted.

REFERENCES

[1] Gause, G. F. 1971. *The Struggle for Existence*. New York: Dover Publications, Inc. First published in 1934 by The Williams & Wilkins Company and available completely on the world wide web at `http://www.ggause.com/Contgau.htm`. Accessed 31 March 2008.

[2] Hutchinson, G. Evelyn. *Introduction to Population Ecology*. New Haven CT: Yale University Press.

[3] Kingsland, Sharon. 1982. The Refractory Model: The Logistic Curve and the History of Population Ecology. *The Quarterly Review of Biology*. 57: 29-52.