www.simiode.org
SIMIODE
A SYSTEMIC INITIATIVE FOR MODELING INVESTIGATIONS
& OPPORTUNITIES WITH DIFFERENTIAL EQUATIONS

STUDENT VERSION
LOGISTIC POPULATION GROWTH MODELING

Brian Winkel Director SIMIODE
Cornwall NY USA

STATEMENT

We offer modeling opportunities in which (1) an artificial data set is given and a model is required and (2) several data sets from an historical protozoan study in the Soviet Union in the 1930's form the basis of the modeling and data. Several different approaches for estimating parameters are offered and the results from these approaches will be compared with each other as well as against the data itself.

First, we examine several elementary growth models, insufficient for our purposes, as background. 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 of the equation with respect to t to obtain $y(t) = k * t + c$, where the initial condition $y(0) = y_0$ determines c, so that we have a complete solution, $y(t) = k * t + y_0$.

A more interesting, and perhaps more applicable, 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

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

The change in (2) in which the rate of change of one quantity is proportional to the amount of said quantity is called *exponential growth* when $k > 0$ (and *exponential decay* when $k < 0$). The parameter k is referred to as the *rate constant*. 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 interest rate in decimal form for 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 these various approaches serve well to estimate k:

1. linearizing 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 (2) to the data.

We seek more appropriate models for growth in biological situations and spread of technology and this demands that we do a better job of modeling the reality of finite resources. This leads to the limited growth model embodied by the *logistic model*.

We can study the logistic differential equation (3) as a model for many phenomena to include the spread of information and technology, the spread of disease, and the growth in various species, with the latter using data from the ecology and population biology literature. Here, as a rich set of examples, we consider a set of microbial data from G. F. Gause's summary publication in his 1934 book *The Struggle for Existence* [1] and classic article [4] of two years earlier, as well as other works [2, 3] which summarize studies conducted in the Soviet Union in the early 1930's on separate and combined populations of several species of paramecia and yeast. These studies are preliminary efforts for population studies on combined yeast populations in the preparation of vodka, a very important commodity in the Soviet Union.

Limited Growth Population Model

In Table 1 (with a plot of the same data in Figure 1) we have population data of a standard sort, i.e. it does not actually grow exponentially. There is a limit to growth. Often this is a natural limit, such as region or resource limitations, geometry of the environment, or borders. To envision this just consider growing microorganisms in a Petri dish which has only a finite space.

Conjecturing a growth model for Table 1 data

Take a moment to consider how we could modify the differential equation (1) for exponential growth so as to actually model growth of data such as found in Table 1 that would level off as illustrated in Figure 1.

1) 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 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? Scrutinize your conjecture and talk about other models being proposed. How reasonable is each model really and what are the good and bad notions for each?

Time	Population Size	Time	Population Size
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 for modeling activity.

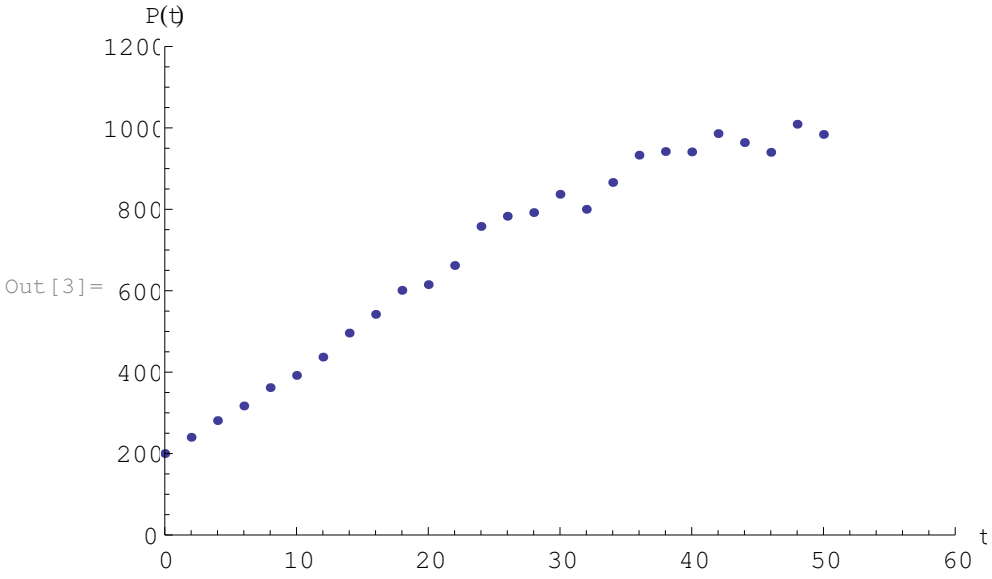

Figure 1. The plot of the logistic-like population data in Table 1.

Now, while not ignoring your efforts, we are going to turn 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.

The differential equation (3) is referred to as the *logistic differential equation*. It is studied in many differential equations courses and is used in such areas as modeling limited growth population biology as well as the study of the spread of disease, technologies, and information:

$$y'(t) = ry(t)\frac{K - y(t)}{K}, \qquad y(0) = y_0\,. \tag{3}$$

2) Take some time to give the logistic model (3) the scrutiny you and your colleagues gave to each other's models.

   Play some "what if" games with it.

   Test it and challenge it.

   Write down your thoughts.

In different contexts (3) could model population or perhaps just the percent of a given population that has heard a specific rumor or acquired knowledge or use of a specific technology. The equation could model the number of a given population that have a disease. The parameter r represents the intrinsic growth rate (in population models r actually is birth rate less death rate) with K as the limiting value of $y(t)$, where $y(t)$ is the variable we are tracking as a function of time, t. K is called the *carrying capacity*. The latter term comes from ecology and refers to the number of the given species that the environment can support or carry. For aneExcellent history of the logistic curve in the history of population biology see [5, 6].

Now suppose we have some artificial data as offered in Table 1. In further work [8] considers the logistic model with actual data from [1] is studied.

3) Upon plotting the data (see Figure 1) and developing a model, presumably (3), based on the sigmoidal or "S" shape of the solution of (3), how might we confirm or deny that (3) is a good model for our data? Get serious, think hard on this.

Throughout this early discussion we may never actually solve (3) by hand - although one could do so using the separation of variables and partial fraction strategies described elsewhere. When we do need an analytic solution we rely on *Mathematica*'s symbolic solving command `DSolve` or its numerical solving command `NDSolve`. Of course, other softeare, e.g., Maple, SAGE, Desmos, etc. are suitable.

In assessing how good a model we have developed in (3) we might consider determining criteria for estimating the parameters. This means finding the parameters which when we substitute them in (3) and solve the differential equation will permit us to compare how our model predicts the actual data in Table 1. So how might we estimate our parameters r and K?

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.

Developing strategies for estimating parameters

Here are several approaches that students usually suggest.

i) Can we estimate the parameters by transforming the model in (3) and then minimize the sum of square differences between transformed data and model to ascertain 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 $b = \frac{r}{K}$?

ii) Can we estimate the parameters by transforming the model in (3) and then minimize the sum of square differences between transformed data and model to ascertain 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 $b = \frac{r}{K}$?

iii) Can we estimate the parameters by analytically solving (3) and then minimize the sum of square differences between data and analytic model to ascertain which parameters yield the least sum of square differences, i.e. nonlinear fit of the analytic solution to the data at the time observations.

iv) Can we estimate the parameters by numerically solving (3) for a wide variety of parameters, minimize the 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 least sum of square differences?

Approach (iv) is a bit strange and rather computationally intense, but it is still doable, namely try lots, and we mean LOTS, of parameter sets r and K in a reasonable vicinity of what you believe might be the best parameters to see if you can get systematically lucky and get a pretty good estimate. We spend considerable time on this approach in [8] because approach (iv) could be useful in the event we cannot analytically solve the differential equation model as we can in this case.

However, let us concentrate on the first two approaches (i) and (ii). Even inside approaches (i) and (ii) 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 (i-a) and (i-b) as well as (ii-a) and (ii-b) with (i) denoting the plain first difference approach and (ii) denoting the symmetric first difference approach. We will then address approach (iii).

In all approaches we shall use the same measure of how good the method is, namely the sum of square errors, i.e. the sum of square of the differences between the value of the solution to (3) evaluated at the time value for each data point and the actual corresponding population value at that time. This would look like:

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

where $\hat{y}(t_i, r, K)$ is the value of our solution to (3) with parameters (r, K) at time t_i (our model), P_i is the observed value of the population at time t_i from our data set for $i = 1, \ldots, n$, and n is the size of our data set. Note that $SSE(r, K)$ has all numbers and is a function or r and K only, while we seek to minimize $SSE(r, K)$.

Parameter estimation opportunities

We offer an opportunity and some guidance in pursuing parameter estimations per the previous section.

4) a) Use approach (i-a) and (i-b) to estimate parameters r and K with the Trendline feature in Excel to fit 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 construct a `ListPlot` of that data, calling the plot dataPlot. Then solve the differential equation and identify (or "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]]
```

Your solution should look like this:

$$ys(t) = \frac{K y_0 e^{rt}}{K + y_0 e^{rt} - y_0}\,. \tag{4}$$

Now after substituting $y_0 = 200$ (see Table 1) and the parameter values for r and K obtained using Excel's Trendline feature into (4) we have a model that we can compare to our data. Plot the data and the model and compare. Discuss your results from (i-a) and (i-b) approaches.

b) In a manner similar to approaches (i-a) and (i-b), except that we transform the data into a linear form, use approach (ii-a) and (ii-b) to estimate parameters r and K using Trendline in Excel to fit a linear function to the transformed data. Use your parameters in each case and return to *Mathematica* to perform a plot of your model over the data and comment on how good this approach is. You might also comment here after performing other analyses in this scenario.

c) Finally, directly estimate r and K by first evaluating the solution in (4) at each time value $t = 0, 2, 4, \ldots, 50$ along with $y_0 = 200$ and form the sum of square errors (5):

$$SSE(r, K) = \sum_{i=1}^{n} (\hat{y}(t_i, r, K) - P_i)^2\,. \tag{5}$$

We can now apply Mathematica's `FindMinimum` command to this sum of square errors to find directly r and K which make the solution to (4) with these r and K values in our best fitting model, i.e. we minimize the sum of square errors, $SSE(r, K)$.

Take these values of r and K and substitute them into (4) along with $y_0 = 200$ and we have a fully developed model which we can plot (call it SSEModelPlot) and compare this plot to our data plot by displaying them together using Mathematica's command `Show`.

Discuss your result and compare it with those from (i-a) and (i-b) as well as with those from (ii-a) and (ii-b) approaches above. What would you use to compare the models?

We do not address approach (iv) here.

Single population data from a classic multiple population study

We apply these approaches to parameter estimation, using logistic growth models for competing species or strains of one species and historical data from G. F. Gause's work [1] and [2], based on studies conducted in the Soviet Union of the 1930's, for studying paramecia and yeast populations.

We point out that Gause was interested in modeling yeast growth (separate species and combined populations) with respect to the production of vodka in state institutions of the Soviet Union. In [8] we discuss his logistic problem for competition, determining the (r, K) parameters in separate population models for two paramecia species he initially studied. Gause was interested in determining what he called the "coefficients for the struggle for existence" (α and β) which model the intensity of the influence of one species upon the other's carrying capacity, all using his data and variations of these several approaches.

In (6) and (7) we have the classical competition model found in [5], for example, in which two populations (different species or strains of the same species), N_1 and N_2, with respective growth rates, r_1 and r_2, and carrying capacities, K_1 and K_2, are in the same environment. This is an example of a system of more than one differential equation which we will study.

The "coefficients for the struggle for existence" (α and β) found in (6) and (7) as Gause called them were of great interest to Gause. What he did was to model each species separately (that is exactly what we investigate here) to discover their r and K values and then he used (6) and (7) to estimate the parameters α and β. In [8] we give extensive coverage of the two population model, but here we confine ourselves to each population separately to offer two data sets for parameter estimations of the respective r's and K's.

In Table 2 we provide the data that Gause obtained from his observations for each population when alone in its environment and we seek to estimate the respective parameter sets, r_1 and K_1 and then r_2 and K_2, for two separate populations in the two separate logistic situations (6) and (7).

$$\frac{dN_1}{dt} = r_1 N_1 \frac{K_1 - N_1}{K_1} \tag{6}$$

$$\frac{dN_2}{dt} = r_2 N_2 \frac{K_2 - N_2}{K_2} \tag{7}$$

We need the data for the species in isolation in order to determine their respective r and K values. Table 2 presents Gause's data of two combined "runs" for each species, in isolation and in mixed population.

5) Estimate the respective parameter sets, r_1 and K_1 and then r_2 and K_2, for Gause's two separate populations (*Saccharomyces cerevisiae* and *Schizosacchaomyces kefir*) in the two separate logistic situations (6) and (7) using the appropriate data in Table 2 which shows the results of several runs of data collection in the puore or unmixed population columns.

   We find Gause's estimates for r and K for each of the two population models in his published works and we include their values in Table 3 for you to compare your work with his. That is what we concentrate on now, leaving the estimates of α and β in (6) and (7) for [8].

Age in hours	*Saccharomyces* Volume of yeast	Mixed Population Volume of yeast	*Schizosaccharomyces* Volume of yeast	Mixed Population Volume of yeast
6	0.37	0.375	-	0.291
16	8.87	3.99	1.00	0.98
24	10.66	4.69	-	1.47
29	12.50	6.15	1.70	1.46
40	13.27	-	-	-
48	12.87	7.27	2.73	1.71
53	12.70	8.30	-	1.84
72	-	-	4.87	-
93	-	-	5.67	-
117	-	-	5.80	-
141	-	-	5.83	-
7.5	1.63	0.923	-	0.371
15.0	6.20	3.082	1.27	0.630
24.0	10.97	5.780	-	1.220
31.5	12.60	9.910	2.33	1.112
33.0	12.90	9.470	-	1.225
44.0	12.77	10.570	-	1.102
51.5	12.90	9.883	4.56	0.961

Table 2. The growth of the yeast volume and the number of cells in pure or unmixed cultures of *Saccharomyces cerevisiae* (column 1), *Schizosaccaromyces kefir* (column 3) and in the mixed population of these species (column 2 and 4 respectively). [4, p. 395]

	Your Analysis		Gause's Analysis	
Species	r	K	r	K
Saccharomyces			0.21827	13.0
Schizosaccharomyces			0.06069	5.8

Table 3. Estimates of Gause [1, p. 78] for r and K values for separate paramecia populations.

6) Using your parameters plot the data and the solution to each of (6) and (7) on the same respective axes and comment on how well your model fits the data.

7) Using Gause's parameters found in Table 3 plot the data and the solution to each of (6) and (7) on the same respective axes and comment on how good Gause's model fits the data.

8) Compare your model with Gause's. You should know that he did all of his calculations by hand using a ruler to estimate slopes from the data and then he used these slopes to advance his

"solutions" for his differential equations.

9) Go back to your original model or take one from a classmate and perform an analysis similar to that which we have done here to determine just how good your model is. Look for supporting evidence that the model is good and also examine it for concerns you might develop in your process.

COMMENTS

In this scenario we go back to the basics of constant growth and exponential growth to help the student transition to limited or logistic growth. We use the notion of limited population growth to motivate the introduction to the logistic equation. We take time in class to have students come up with limiting factors before introducing them to the classic logistic differential equation (3). We discuss how we might estimate the parameters.

The paper [8] goes into great depth in terms of analysis and parameter estimation for single and multiple population models based on these data and the logistic model. We offer a *Mathematica notebook* (1-18-Mma-T-LogisticProModel-TeacherVersion.nb and .pdf) in which the analyses for activities (4) - (8) are offered. Activities (1) - (3) and (9) are for teacher consideration and you will find that results vary, but if students are truthful the reflection on these activities will give students great insight into the modeling process.

A paper with other sources of data collection as well as collected data which can be modeled by the logistic equation can be found in [9].

In a final section of the *Mathematica notebook* (1-18-Mma-T-LogisticProModel-TeacherVersion.nb and .pdf) we offer data for the mean number of *Paramecium caudatum*, in a culture at time t in days. This data is found in Table 3, *P. Caudatum*, Mean Number of individuals per 0.5 c.c.[1, p. 144]. We use and compare the strategies developed here. This provides another data set on which students can develop their modeling skills.

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 https://asantos.webs.ull.es/The Struggle for Existence.pdf. Accessed 17 February 2020.

[2] Gause, G. F., O. K. Nastukova, and W. W. Alpatov. 1934. The Influence of Biologically Conditioned Media on the Growth of a Mixed Population of *Paramecium cadatum* and *P. aureliax*. *Journal of Animal Ecology*. 3(2): 222-230.

[3] Gause, G. F. and W. W. Alpatov. 1931. Die logistische Kurve von Verhulst-Pearl und ihre Anwendung im Gebiet der quantitativen Biologie. *Biol. Zentralbl.* 51: 1-14.

[4] Gause, G. F. 1932. Experimental Studies on the Struggle for Existence. *Journal of Experimental Biology*. 9(4): 389-402. Available on-line at http://jeb.biologists.org/cgi/reprint/9/4/389.pdf. Accessed 3 January 2010.

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

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

[7] Winkel, B. J. 2011. Parameter Estimates in Differential Equation Models for Chemical Kinetics. *International Journal of Mathematical Education in Science and Technology*. 42(1): 37-51.

[8] Winkel, B. J. 2011. Parameter Estimates in Differential Equation Models for Population Growth. *PRIMUS*. 21(2): 101-129.

[9] Winkel, B. J. 2012. Sourcing for Parameter Estimation and Study of Logistic Differential Equation. *International Journal of Mathematical Education in Science and Technology*. 43(1): 67-83.