Technological Forecasting and Social Change 61(3):247–271, 1999.

A Primer on Logistic Growth and Substitution: The Mathematics of the Loglet Lab Software

PERRIN S. MEYER, JASON W. YUNG[1] and JESSE H. AUSUBEL[2]

ABSTRACT

This article describes the mathematics underlying the Loglet Lab software package for loglet analysis. "Loglet analysis" refers to the decomposition of growth and diffusion into S-shaped logistic components, roughly analogous to wavelet analysis, popular for signal processing and compression. The term "loglet" joins "logistic" and "wavelet." Loglet analysis comprises two models: the first is the component logistic model, in which autonomous systems exhibit logistic growth. The second is the logistic substitution model, which models the effects of competitions within a market. An appendix describes the current status of the software.

1. Introduction

We are all accustomed to the idea of growth to a limit, for example, the number of people becoming ill in an epidemic. In fact, observers have recorded thousands of examples of such S-shaped growth in settings as diverse as animal populations [1], energy and transport infrastructures[2, 3, 4], language acquisition[5], and technological performance[6, 7, 8]. Often the measured quantity (population of a species, height of a plant, power of an engine) grows exponentially at the outset. However, natural systems cannot sustain exponential growth indefinitely. Rather, negative feedback mechanisms or signals from the environment slow the growth, producing the **S-shaped curve**. Thus, for a single growth process, a single sigmoidal curve is often a useful model.

However, many systems exhibit complex growth, with multiple processes occurring sequentially or simultaneously. Kindleberger [9], in an economic history of the world since 1500, wrote: "In the real world there are many wiggles, speedups, and setbacks, new S-curves growing out of old, separate curves for different sectors and regions of a national economy, all of which present difficulties when an attempt is made to aggregate them on a weighted basis." Many such phenomena, it turns out, can be described elegantly with a simple mathematical model.

Humans create technologies, some of which are selected and diffused into society, much like servers on a network creating waves of packets and sending them out onto

[1]PERRIN S. MEYER and JASON W. YUNG are Research Assistants at the Program for the Human Environment, The Rockefeller University, New York, NY, USA.

[2]JESSE H. AUSUBEL is Director at he Program for the Human Environment, The Rockefeller University, New York, NY., USA.

Address Correspondence to: Perrin Meyer, The Rockefeller University, 1230 York Ave., Box 234, New York, NY., 10021. Phone: 212–327–7842. E–mail: `perrin@phe.rockefeller.edu` Website: `http://phe.rockefeller.edu`

the Internet. Aggregating all these wavelets [3] creates much of the apparent complexity we observe; in this case, the problem of decoding complexity is essentially a problem of deconvolution. Generally speaking, the human eye is not well-suited to perform this task with a reasonable degree of precision. The Loglet Lab software package was designed to assist us in this endeavor.

We propose the development of **loglet analysis** for the the **analysis**, **decomposition**, and **prediction** of **complex growth processes**. The term "loglet", coined at The Rockefeller University in 1994, joins "logistic" and "wavelet." Two main objectives of loglet analysis are to analyze existing time-series growth data sets in order to decompose the growth process into sub-processes and to elucidate information on carrying capacities and other aspects ("top-down" approach); and to analyze individual sub-processes in order to determine macro or envelope system behavior ("bottom-up" approach). At the heart of loglet analysis is the three-parameter S-shaped **logistic growth model**. The logistic is attractive for modeling S-shaped growth because it is a parsimonious model where the three parameters have clear, physical interpretations.

The **Loglet Lab** software package allows users to perform loglet analysis on any suitable time-series data set. The user interface is easy and informative for the casual user and the common case of a single logistic in isolation; at the same time, Loglet Lab has an advanced fitting engine to analyze complex data and compound phenomena. This document provides the background for performing logistic analysis. The following sections describe the mathematics behind the Loglet Lab software package. For instructions on using the Loglet Lab software package, consult the "Loglet Lab Tutorial" [10].

2. Simple Growth Models

The mathematical models used in this article are based on Ordinary Differential Equations (ODEs). Physicists first used ODEs to model the trajectories of moving objects. When applied to population or technologies, they describe continuous "trajectories" of growth or decline through time. Although populations of humans and many technological variables grow and decline in discrete numbers, continuous models are often used for simplicity when modeling large aggregates. Montroll [11] makes the connection between physical and population "trajectories" clear by proposing "laws of social dynamics" based on Newton's laws of mechanics.

The exponential growth of multiplying organisms is represented by a simple and widely used model that increases without bounds or limits as Figure 1 illustrates. In mathematical terminology, the growth rate of a population $P(t)$ is proportional to the population. The growth rate at time t is defined as the derivative $\frac{dP(t)}{dt}$. The exponential growth model written as a differential equation

$$\frac{dP(t)}{dt} = \alpha P(t) \qquad \text{Exponential Growth Model} \tag{1}$$

[3] In this article, we use the wavelet analogy to motivate our formulation of Loglet analysis. The mathematical equivalence of what wavelet analysis currently means in the digital signal processing community and our model is a topic of our active research. In traditional wavelet analysis, the basis function must be orthogonal, and the logistic model is not. We suspect that our model could fit into the framework of non-orthogonal wavelet theory. However, results from this theory are not necessary for any of the methods presented in this article.

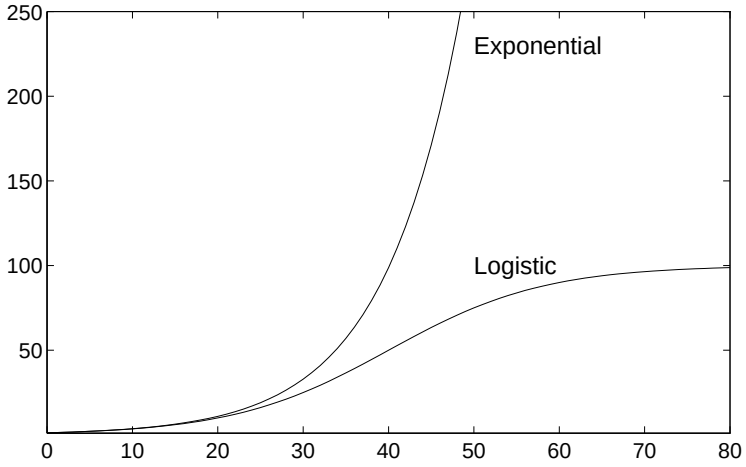

FIGURE 1. Comparison of exponential and logistic growth.

can be solved by introducing e (the base of the natural logarithm, approximately $2.71\ldots$). The familiar solution to (1) is

$$P(t) = \beta\, e^{\alpha\, t} \tag{2}$$

where α is the growth rate constant and β is the initial population $P(0)$. α is often expressed in percent. An α with a value of 0.02 is equivalent to the statement "the population was growing continuously at 2% per year." Although many populations grow exponentially for a time, no bounded system can sustain exponential growth indefinitely unless the parameters or boundaries of the system are changed.

Because few, if any, systems are permanently unbounded and sustain exponential growth, equation(1) must be modified with a limit or a carrying capacity that gives it the more realistic sigmoidal shape of the lower curve in Figure 1. The most widely used modification of the exponential growth model is the **logistic**. It was introduced by Verhulst in 1838 but popularized in mathematical biology by Lotka [12] in the 1920's, as Kingsland [13] writes in her comprehensive history of such models in population ecology.

The logistic equation begins with the $P(t)$ and α of the exponential but adds a "negative feedback" term $\left(1 - \frac{P(t)}{\boldsymbol{\kappa}}\right)$ that slows the growth rate of a population as the limit $\boldsymbol{\kappa}$ is approached:

$$\frac{dP(t)}{dt} = \alpha P(t) \underbrace{\left(1 - \frac{P(t)}{\boldsymbol{\kappa}}\right)}_{\text{feedback term}} \tag{3}$$

Notice that the feedback term $\left(1 - \frac{P(t)}{\boldsymbol{\kappa}}\right)$ is close to 1 when $P(t) \ll \boldsymbol{\kappa}$ and approaches zero as $P(t) \to \boldsymbol{\kappa}$. Thus, the growth rate begins exponentially but then decreases to zero as the population $P(t)$ approaches the limit $\boldsymbol{\kappa}$, producing an S-shaped (sigmoidal) growth trajectory.

It is possible to solve the logistic differential equation (3) to find an analytic (algebraic) solution. Often, more complicated differential equations do not have

analytic solutions, and must either be simplified or solved numerically [14]. The solution to the logistic differential equation (3) is:

$$P(t) = \frac{\boldsymbol{\kappa}}{1 + \exp\left(-\alpha(t - \beta)\right)} \tag{4}$$

Equation (4) produces the familiar S-shaped curve. Note that three parameters are needed to fully specify the curve, α, β, and $\boldsymbol{\kappa}$.

The growth rate parameter α specifies the "width" or "steepness" of the sigmoidal curve. It is often helpful to replace α with a variable that specifies the time required for the trajectory to grow from 10% to 90% of the limit $\boldsymbol{\kappa}$, a period which we call the **characteristic duration**, or $\mathbf{\Delta t}$. Through simple algebra, the characteristic duration is related to α by $\mathbf{\Delta t} = \frac{\ln(81)}{\alpha}$. The parameter Δt is usually more useful than α for the analysis of historical time-series data because the units are easier to appreciate. The parameter β specifies the time when the curve reaches $\frac{1}{2}\boldsymbol{\kappa}$, or the **midpoint** of the the growth trajectory, often re-labeled $\mathbf{t_m}$. The parameter $\boldsymbol{\kappa}$, as discussed, is the asymptotic limit that the growth curve approaches, i.e., market niche or carrying capacity. The logistic model is symmetric around the midpoint t_m. Other models describe non-symmetric or skewed growth, e.g., the Gompertz. We will not consider non-symmetric growth in this study. Banks [15] comprehensively surveys growth and diffusion phenomena modeled by ODE's.

We can readily compare the exponential and logistic growth models. In the examples in Figure 1, both models have the same growth rate parameter $\alpha = 0.11$ (approximately 11% per year) and a starting population of 1.22 at $t = 0$. Note that for the first 20 years the exponential and logistic curves are hardly distinguishable, and only diverge significantly after 30 years. But after only 50 years the exponential curve has exploded off the chart while the logistic is stabilizing near the carrying capacity $\boldsymbol{\kappa}$, in this case 100. The characteristic duration of this logistic, the time needed for the population to grow from 10% to 90% of $\boldsymbol{\kappa}$, is 50 years, with a midpoint $t_m = 40$.

Figure 2 illustrates the rewarding fit of the logistic to the multiplication of bacteria consuming sugar and minerals in a closed petri dish and especially to their stagnation when the food runs out or they befoul their environment. Here the carrying capacity $\boldsymbol{\kappa}$ is limited by available space (which equates with food). As the bacteria exhaust the nutritious area of the dish, the growth rate slows, producing the S-shaped logistic growth trajectory.

The three parameters κ,Δt, and t_m define the parameterization of the logistic model used as the basic building block for Loglet analysis

$$N(t) = \frac{\kappa}{1 + \exp\left[-\frac{\ln(81)}{\Delta t}(t - t_m)\right]}. \tag{5}$$

The simple logistic model is useful in part because the parameters obtained by fitting the model to data can be easily compared across many different systems. These parameters also help in formulating complex models, for which they provide a frame of reference and first-order guesses for the possible time-scales and magnitudes of component variables.

2.1. **Visualization of the logistic model.** Usually we first visualize logistic growth by simply plotting data on an absolute and linear scale. A change of variables that normalizes a logistic curve renders it a straight line. This view is known

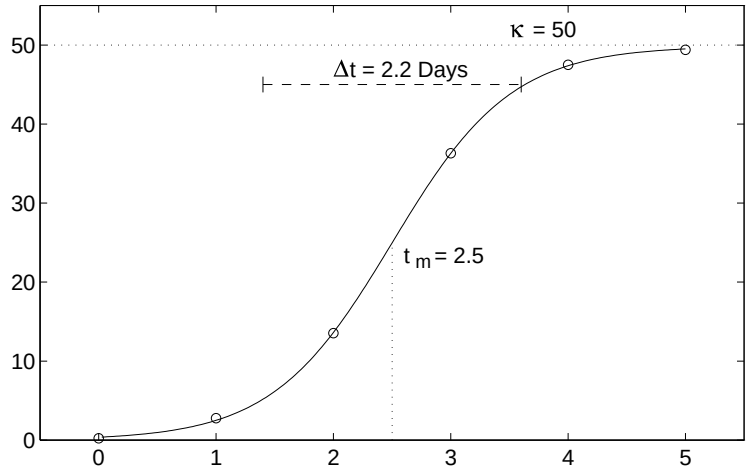

FIGURE 2. Growth of a bacteria colony fitted to a logistic curve. Source of data: [16].

as the **Fisher-Pry**[4] **Transform**:

$$FP(t) = \left(\frac{F(t)}{1 - F(t)}\right), \text{ where } F(t) = \frac{N(t)}{\kappa} \tag{6}$$

Note that

$$\ln(FP(t)) = \frac{\ln(81)}{\Delta t}(t - t_m) \tag{7}$$

so if $FP(t)$ is plotted on a semi–logarithmic scale, the S-shaped logistic is rendered linear. Figure 3 shows the Fisher-Pry transform of the bacteria example in figure 2. We observe that the time in which the value is between 10^{-1} and 10^{1} is equal to Δt, and the time at 10^{0} is the point of inflection (t_m). On the right axis we label the corresponding percent of saturation ($100 * F$) at each order of magnitude from 10^{-2} to 10^{2} rounded to the nearest percent. Because the Fisher-Pry transform normalizes each curve to the carrying capacity κ, more than one logistic can be plotted on the same chart for comparison. As we will see, this becomes useful when we analyze more complex growth behaviors.

3. The Component Logistic Model

Many growth and diffusion processes consist of several subprocesses. First, let us model a system that experiences growth in two discrete phases. Then, we will extend this model to an arbitrary number of phases.

Systems with two growth phases we have termed "**Bi-logistic"** [7]. In this model, growth is the sum of two discrete "wavelets" , each of which is a three-parameter logistic:

$$N(t) = N_1(t) + N_2(t), \tag{8}$$

[4]Named after the authors who wrote a paper that popularized this technique [6], the method first appeared in a paper written by Edwin B. Wilson in 1925 [17].

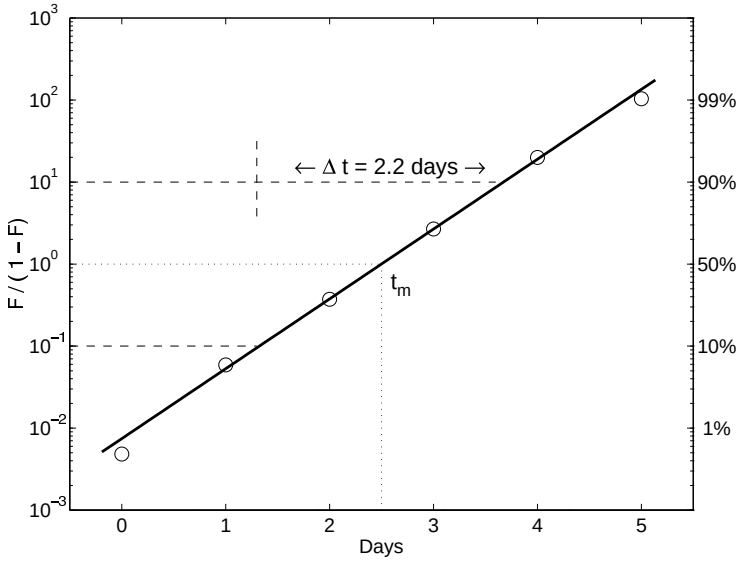

FIGURE 3. Logistic growth of a bacteria colony plotted using the Fisher–Pry transform that renders the logistic linear.

where

$$N_1(t) = \frac{\kappa_1}{1 + \exp\left[-\frac{\ln(81)}{\Delta t_1}(t - t_{m1})\right]}$$

$$N_2(t) = \frac{\kappa_2}{1 + \exp\left[-\frac{\ln(81)}{\Delta t_2}(t - t_{m2})\right]}$$

We can examine system-level behavior (i.e., $N(t)$), or we can decompose the model and examine the behavior of the discrete components (either $N_1(t)$ or $N_2(t)$). In fact, we can plot $N_1(t)$ and $N_2(t)$ on the same axes, and moreover we can compare two disparate loglets by normalizing them with the Fisher-Pry transform.

3.1. **Taxonomy of bi-logistic curves.** Wavelets often overlap in time. Depending on the order and magnitude of the overlap, the aggregate curve can take on a wide range of appearances. Figure 4 shows a taxonomy of bi-logistic processes, with the Fisher-Pry transform of the two component logistics on the right.

Panel A is an example of a "sequential" bi-logistic; the second pulse does not start growing until the first pulse has nearly reached its saturation level κ_1. This shape bi-logistic characterizes a system which pauses between growth phases.

Panel B is an example of a "superposed" bi-logistic, where the second pulse begins growing when the first pulse has reached about 50% of saturation. This bi-logistic growth model characterizes systems that contain two processes of a similar nature growing concurrently except for a displacement in the midpoints of the curves.

Panel C shows a "converging" bi-logistic, where a first wavelet is joined by a second faster, steeper wavelet; the two pulses culminate at about the same time. Often a late adopter of a technology, having learned from the experiences of an early adopter, will advance faster, resulting in a smaller Δt.

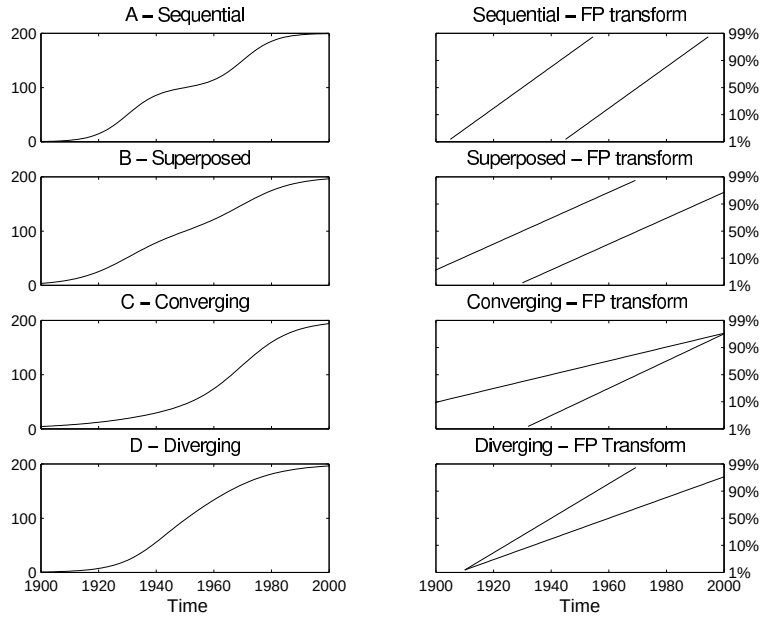

FIGURE 4. A taxonomy of bi-logistic processes growing through time to a notional limit of 200 units. Fisher-Pry decompositions are shown in the right column.

Panel D shows a "diverging" bi-logistic, where two logistic growth processes begin at the same time but grow with different rates and carrying capacities defined from the start.

The panels show the merits of Loglet analysis. While curve A looks logistic, curve B hardly appears S-shaped. Curves C and D are S-shaped, but asymmetric, so they do not appear to be logistic. Yet all four curves are made up of logistic components.

3.2. **Generalization of the Bi-logistic model.** Now we generalize the bi-logistic model to a *multi*-logistic model, where growth is the sum of n simple logistics:

$$\mathbf{N}(t) = \sum_{i=1}^{n} \mathbf{N_i}(t), \tag{9}$$

where

$$\mathbf{N_i}(t) = \frac{\kappa_i}{1 + \exp\left[-\frac{\ln(81)}{\Delta t_i}(t - t_{mi})\right]}. \tag{10}$$

Figure 5 shows a Loglet analysis of a hypothetical data set fitted with the sum of five component logistics (shown in the box in the upper right hand corner). Here again, apparently complex behavior reduces to the sum of logistic wavelets. Note that "growth" processes also include processes of decline; in our model, this occurs when $\Delta t < 0$.

3.3. **Implementation in Loglet Lab.** The Loglet Lab software package allows users to fit several logistic pulses to a series, elucidating clues to the components

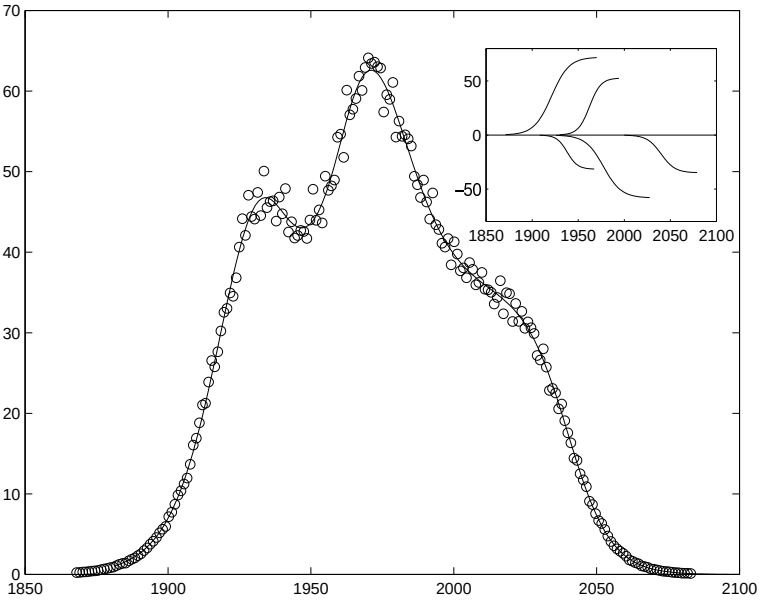

FIGURE 5. A loglet analysis of a data set composed of five logistics.

that form the wiggles, speedups, and set backs that Kindleberger described so well. Users are prompted to provide the fitting engine with a bare amount of information (i.e., how many components to use); the fitting engine, as its name would imply, does the mathematical calculations to obtain the parameters which optimally describe the logistic, and plots the resulting curve. Once the fit has been made, users can access the various visualizations which we described above.

The following section will describe the mathematics of the fitting engine and the mathematical motivations behind analyzing systems with logistic growth.

4. The Mathematics of Loglet Analysis

This section introduces the algorithms implemented in Loglet Lab for fitting logistics with arbitrary numbers of components to a data set. We include examples using hypothetical and historical data, as well as descriptions of the corresponding operations in Loglet Lab.

Because growth over time of either organisms or the diffusion of some other variable such as number of cars or length of canals is the usual subject of loglet analysis, vector notation is convenient to represent the data values, the parameters, and the fitted Loglet model.

4.1. **Definitions and notation.**

Consider the two-dimensional space in which our data set D exists. If there are m data points, then we define D as

$$D = \{(t_1, d_1), \ldots, (t_i, d_i), \ldots, (t_m, d_m)\}$$

where t_i usually represents time, while d_i represents the growing variable (e.g., number of organisms, percent of saturation).

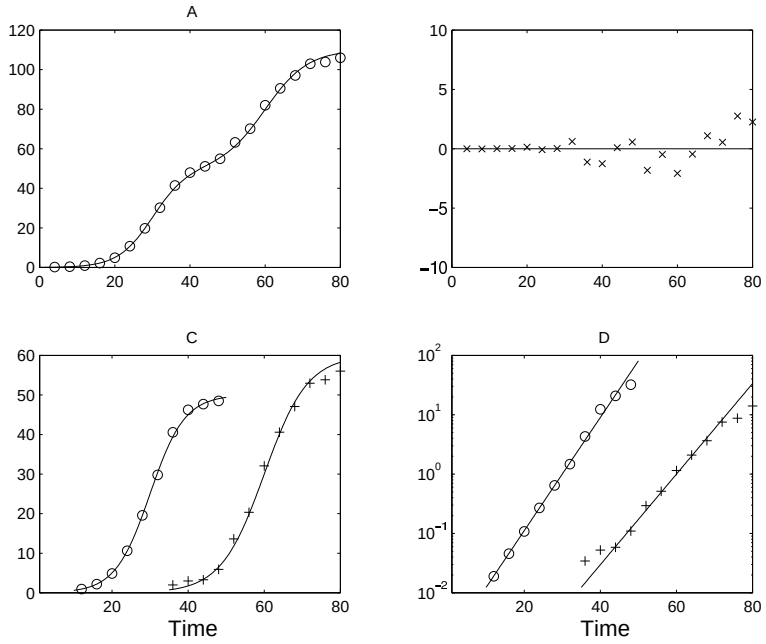

FIGURE 6. A hypothetical bi-logistic data set(A), residuals of the fitted curve(B), and decompositions in raw form (C) and with the Fisher-Pry transform applied (D).

Suppose we want to fit a logistic curve of n components to the model. Then we will require $3n$ parameters, represented as a $n \times 3$ matrix $\mathbf{P}$, where the ith row describes the ith component:

$$\mathbf{P} = \begin{bmatrix} \Delta t_1 & \kappa_1 & t_{m1} \\ & \vdots & \\ \Delta t_n & \kappa_n & t_{mn} \end{bmatrix}$$

Thus a loglet can be alternatively specified by

$$\mathbf{N}(t, \mathbf{P}) = \sum_{i=1}^{n} \frac{\mathbf{P}_{i2}}{1 + \exp\left[-\frac{\ln(81)}{\mathbf{P}_{i1}}(t - \mathbf{P}_{i3})\right]}$$

Figure 6A shows a hypothetical data set (the circles) and a fitted loglet with $n = 2$ and

$$\mathbf{P} = \begin{bmatrix} 20 & 50 & 30 \\ 25 & 60 & 60 \end{bmatrix}.$$

Gaussian noise was added to the data so the data to show non-zero residuals.

4.2. **Decomposition.** Now consider the set of n logistic components $C = \{c_1, \ldots, c_n\}$, where

$$c_i = N_i(x_i, \mathbf{P}_i)$$

where x_i is an arbitrary subspace of t. (The components c_i are analogous to $N_i(t)$ as discussed in equation (10).) A typical choice for x_i is the interval over which c_i grows from 10% to 90% of its saturation level, namely $(t_{mi} - \Delta t_i, t_{mi} + \Delta t_i)$.

Restricting the domain of c_i gives us a criterion for decomposing the associated data set D into subsets D_j $(1 \leq j \leq n)$ corresponding to each component logistic c_i. A subset D_j contains a point (t_i, d_i) if $t_{mi} - \Delta t_i \leq t_i \leq t_{mi} + \Delta t_i$:

$$D_j = \{(t_i, d_i^*) \mid t_i \in x_i\}.$$

where d_i^* is the *adjusted* value of d_i. The adjustment is subtracting out the "effects" from other components, leaving us with the (approximate) contribution of component c_i to this data point. In other words,

$$d_i^* = d_i - \sum_{j \neq i} \frac{\mathbf{P}_{j2}}{1 + \exp\left[-\frac{\ln(81)}{\mathbf{P}_{j1}}(t - \mathbf{P}_{j3})\right]}$$

In Figure 6C, the hypothetical data set plotted in Figure 6A is decomposed into subsets D_1 (circles) and D_2 (crosses); similarly, the fitted curve is also decomposed into its component logistics. Note that the subsets D_j are not necessarily mutually exclusive in the t domain; in this example, D_1 and D_2 share points with common t-values around $t = 40$. At these times, we can see that there are two concurrent growth processes; in addition, we can also quantify how much of the growth can be attributed to each process.

As we saw in Figure 4, we can apply the Fisher-Pry transform to each component and its corresponding data subset. This is useful because it normalizes each component on a semi-logarithmic scale, allowing for easy comparison when plotted on the same graph. Moreover, it allows the fitting of logistics using linear least squares. The Fisher-Pry transform of components $\mathbf{c}_i$ is

$$FP(c_i) = \frac{\frac{\mathbf{c}_i}{\kappa_i}}{1 - \frac{\mathbf{c}_i}{\kappa_i}}$$

where plotting $\mathbf{x}_i$ vs. $\log(FP(c_i))$ produces a straight line. The component data subsets D_j are transformed in a similar manner:

$$FP(d_i^*) = \frac{\frac{d_i^*}{\kappa_i}}{1 - \frac{d_i^*}{\kappa_i}}$$

Figure 6D shows the Fisher-Pry decomposition of the hypothetical data, while Figure 6B shows the residuals, as discussed later in section 4.4.

4.3. **Growth rates and the "bell" view.** Just as the differential equation (3) reveals the mechanism propelling its integral equation (4), the rates of change of the component logistics provide clues to the mechanisms propelling the composite logistic. Analyzing the rates of change is often useful when yearly or percent per year tabulations are applicable, as in the case of economic data. Recall that the analytic form of the three-parameter logistic is:

$$N(t) = \frac{\kappa}{1 + \exp\left[-\frac{\ln(81)}{\Delta t}(t - t_m)\right]}$$

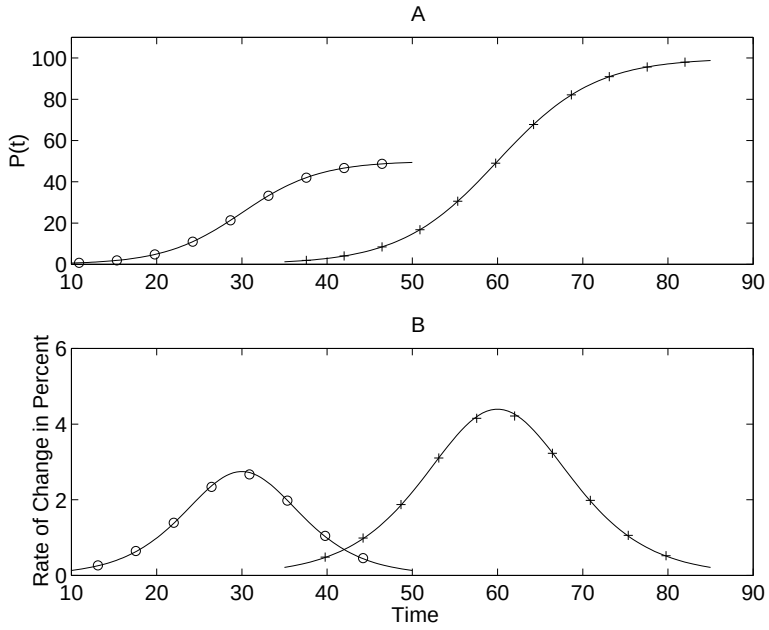

FIGURE 7. Rates of change of the two component logistics (The "Bell View")

The instantaneous rate of growth of the logistic function is given by its derivative with respect to time:

$$\frac{d\,N(t)}{d\,t} = \frac{\frac{\ln(81)}{\Delta t}\ \kappa\ \exp\left(-\frac{\ln(81)}{\Delta t}(t-t_m)\right)}{\left[1+\exp\left(-\frac{\ln(81)}{\Delta t}(t-t_m)\right)\right]^2} \tag{11}$$

Plotting (11) produces a bell-shaped curve similar, but not identical, to the normal distribution function. Naturally, the bell-shaped curve peaks at the midpoint t_m; analytically, this is the point of inflection, and thus it is an extremum of $N(t)$. Panel B of Figure 7 shows the derivative of the component logistics of our test function (Panel A is shown again for comparison purposes).

The rate of change of the data subsets D_j is computed discretely, creating the sets $\overline{D_j}$, in which each point E_i is

$$E_i = \left(\frac{t_i + t_{i+1}}{2} + t_i, \frac{\mathbf{c}_i(t_{i+1}) - \mathbf{c}_i(t_i)}{t_{i+1} - t_i}\right)$$

where $1 \le j \le m-1$. In other words, E_i contains the discrete derivative from (t_i, d_i^*) to (t_{i+1}, d_{i+1}^*).

Loglet Lab can decompose a logistic curve into its discrete components. Each component can be transformed using the Fisher-Pry transform, and its rate of change can be plotted.

4.4. **Residuals.** Residuals are the error, or difference, between the model and the observed data. The residual vector $\mathbf{R} = \{r_1, \ldots, r_m\}$ is defined by

$$r_i = d_i - N\,(t, \mathbf{P}).$$

For the hypothetical data set actual values of the residual vector were plotted in Figure 6B.

We can also calculate residuals as percentage error:

$$r_i = \frac{(d_i - N(t, \mathbf{P}))}{N(t, \mathbf{P})} \times 100.$$

It is crucial to examine the residuals after a fit. When a fit is "good," the residuals are non-uniformly distributed around the zero axis; that is, they appear to be random in magnitude and sign. A substantial or systematic deviation from the zero axis indicates some phenomenon is not being modeled or fitted correctly. An iterative process of fitting loglets to a data set and then examining the residuals is a good way to proceed, unless the errors in the data and shown in the residuals are known to come from other sources (e.g., an economic recession). Loglet Lab provides views of both percentage and raw error.

5. Numerical Methods for Estimating Loglet Parameters from Time-Series Data

The Loglet model is nonlinear. Although there are no direct methods for estimating the parameters for nonlinear models, we can use iterative methods for this purpose. Such methods minimize some function of the residuals.

The standard method for estimating model parameters is the method of least-squares, where the sum of the squares of the residuals is minimized. In our notation, our goal is to

$$\text{vary } \mathbf{P} \text{ such that } \quad \chi^2 = \sum r_i^2 \quad \text{is minimized.}$$

Thus we must set $\mathbf{P}_0$, which holds initial values for $\mathbf{P}$, and iteratively adjust its entries until χ^2 has sufficiently converged to a minimum. Note that we do not have to adjust all the entries of $\mathbf{P}_0$; there may be reason to hold any one of the entries constant. For example, there may be physical constraints to the growth (the size of the petri dish limits the population of a bacteria culture), or time constraints on the midpoint or growth time.

The least-squares method assumes errors are randomly and normally distributed. However, predetermining the error distribution of historical data sets is often hard. Least-squares can still be used, but the parameter value estimates are no longer guaranteed to be correct. In fact, on data sets with outliers, or systematic errors, least-squares regression can produce poor results.

For example, least-squares parameter estimates for logistic functions can overestimate the saturation value (κ). For incomplete S-shaped processes the estimation of κ by OLS depends strongly on small deviations around the midpoint t_m. Thus, when fitting an incomplete S-shaped process in Loglet Lab, it is usually a good idea to try a second fit with the saturation held at, say, 90% of the final value from the first fit and compare the new fit as well as the new residuals. In addition, we have found that using the Fisher-Pry transform to corroborate the fit can help produce useful results, as the FP transform in effect weighs the data points near the beginning and tail of the sigmoid more than near t_m.

5.1. **Weighting and masking.**

In some historical data sets, certain data may be known to be affected by external agents such as war, or there may be some other bias or problem. Variation in the quality or reliability of data provides an

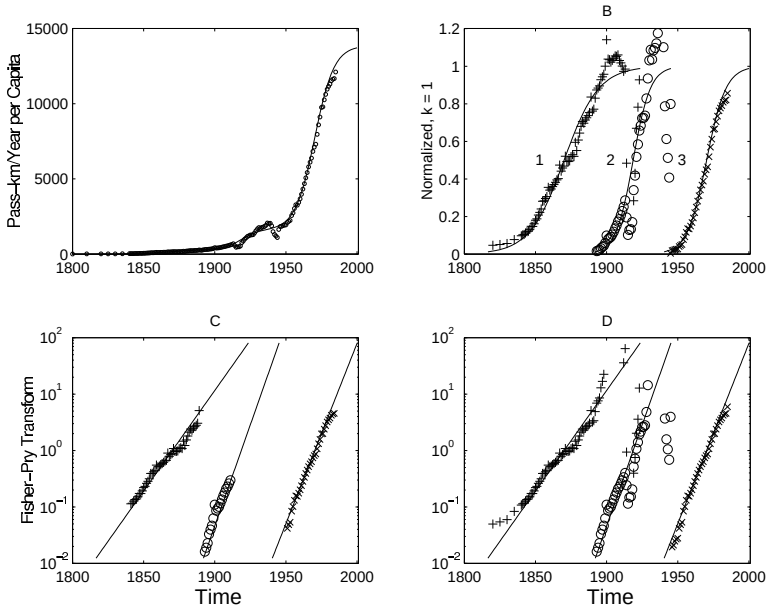

FIGURE 8. Analysis of French Motorized Mobility. Panel B shows the component logistics normalized to their respective κ_i (for scale). Panel C shows the Fisher-Pry transform of the masked data set, while Panel D shows the Fisher-Pry transform of all of the (unmasked) data.

incentive for weighting data sets. This is accomplished by introducing a weight vector of length m $\mathbf{w} = \{w_1, \ldots, w_m\}$ that contains the weight for each data point d_i. Weighted least squares is thus denoted

$$\text{vary } \mathbf{P} \text{ such that } \quad \sum \left(\frac{r_i}{w_i} \right)^2 \quad \text{is minimized}$$

If $\mathbf{w}_i = 1$ for all i, then we have the same model as before.

With historical time-series data, a more accurate analysis may be achieved by focusing on "quiet" periods and excluding unrepresentative data. For example, an analysis of nuclear testing data might be improved by excluding data from the turbulent years around the signing of the Nuclear Test Ban Treaty. Exclusion is accomplished by setting $\mathbf{w}_i = \infty$ for certain values of i. This is sometimes referred to as **"masking"** data, as we are hiding some of the points from the fitting engine.

Loglet Lab accommodates masking of the data. However, it does not allow use of user-specified weight vectors, though this functionality could be added in the future.

5.2. **French Mobility: an example.** This example illustrates logistic analysis of a data set with several components and masking. Figure 8 presents the results of an analysis of historical time-series data of French motorized mobility[5]. Periods of conflict, such as the World Wars and Great Depression cause substantial deviations from normal activity; analysis of long term trends, then, should focus on the years

[5]Source of data: Arnulf Grübler, IIASA, Laxenburg, Austria

outside of these periods. That said, for this analysis, we excluded the data between 1912 and 1950. Accordingly, the weights **w** for the corresponding data points were set to ∞.

We posit logistics driven successively by the diffusion of homes, railroads, and automobiles. Figure 8a shows the data set D (circles) and the estimated fitted curve $N(t, \mathbf{P})$, where

$$\mathbf{P} = \begin{bmatrix} 53 & 322 & 1870 \\ 26 & 1291 & 1918 \\ 29 & 12254 & 1970 \end{bmatrix}.$$

The values in the middle column suggest rail quadrupled French mobility over horses, and autos tenfold over rail. We could posit another loglet driven by aviation and high speed rail.

5.3. **Confidence Intervals on the Estimated Parameters: The Bootstrap.** An important question to ask of a least-squares fit is "How accurate are the estimated parameters for the data?"

In classical statistics, we are accustomed to having at our disposal not only a single value for an estimated parameter, but also a *confidence interval* (CI) within which its "true" mean is expected to lie. To ascertain the statistical errors of estimated parameters, the errors of the underlying data must be known. For example, if we know that the measurement errors for a particular dataset are normally distributed (a common assumption) with a known variance, we can estimate the error of the parameters.

However, for historical data sets, it is often impossible to know the distribution and variance of the errors in the data, and thus impossible to estimate the error in the fit. However, the **bootstrap method** [18] provides a means for recreating and resampling data using Monte Carlo methods. While the effectiveness of the bootstrap method has been known for decades, its implementation requires computational power largely unavailable until recently but now common on PC's.

The bootstrap method uses the residuals from the least squares fit to synthesize data sets. We generate, say, 200 data sets and fit a curve to each set, giving us 200 sets of parameters. By the Central Limit Theorem, we assume the bootstrapped parameter estimates are normally distributed around a sample mean. From these sets we can proceed to compute confidence intervals for the parameters. From the confidence intervals of a parameter, we can form a confidence *region* which contains the set of all curves corresponding to all values of that parameter.

For a data set $\mathbf{D}_0$ with m points, we first estimate the loglet parameters $\mathbf{P}_0$ using the least-squares algorithm described above and calculate the residuals $\mathbf{R}$. Then we synthesize n_{boot} data sets adding $\mathbf{R}_i^*$, a vector containing m residuals chosen at random (with replacement) from $\mathbf{R}$:

$$\mathbf{D}_i^* = N(t, \mathbf{P}_0) + \mathbf{R}_i^*$$

For each $\mathbf{D}_i^*$, the bootstrap parameters $\mathbf{P}_i^*$ are estimated. In Loglet Lab, the default value for n_{boot} is 200, but this number can be varied depending on the number of data. Larger datasets may require more runs for accurate statistics. The estimated parameters from each fit are stored in a three-dimensional matrix $\mathbf{P}_{boot}$.

Because the distribution of the parameters in $\mathbf{P}_{boot}$ is assumed to be normal, the 95% CI for any parameter p_{ij} in $\mathbf{P}_{boot}$ can be computed from the mean μ_{ij} and

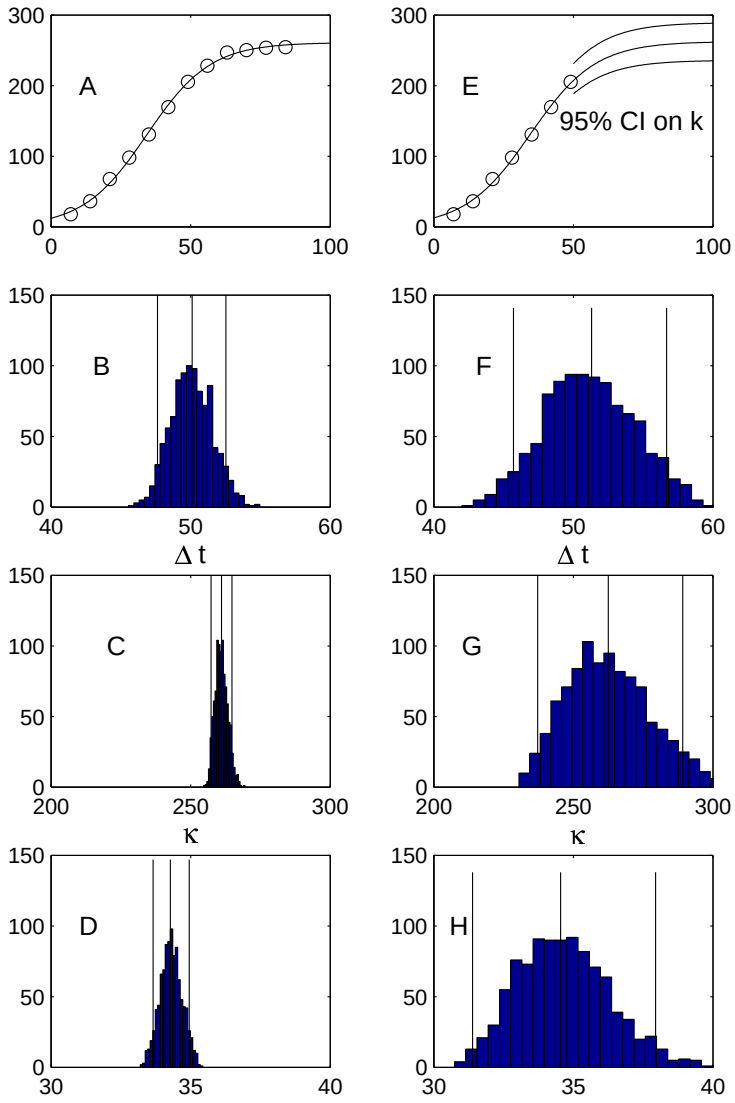

FIGURE 9. 95% bootstrap confidence intervals on the estimated parameters for the growth of a sunflower. The left side (A-D) shows the C.I. using all time series data, the right side (E-H) shows the C.I. obtained by masking the last five data points. Data from Thornton [16].

standard deviation σ_{ij}:

$$95\%\ \text{CI}\ = (\mu_{ij} - 1.96\,\sigma_{ij},\ \mu_{ij} + 1.96\,\sigma_{ij})$$

Figure 9 shows a bootstrap analysis of the Growth of a Sunflower (a "classic" logistic fit, available in the Loglet Lab gallery). Panel 9A shows the sunflower data

fitted with a single logistic, with the parameter values estimated using the least-squares algorithm, $\Delta t = 50$, $\kappa = 261$, and $t_m = 34$. Panels 9B, C, and D show histograms of the distributions of each parameter as determined by 1000 iterations of the bootstrap algorithm described above, along with the mean and 95% C.I. marked by the solid lines.

To show how the completeness of a data set influences the confidence interval, Panel 9E fits a single logistic to the same data, but now with the last five data points masked. The upper and lower solid lines show the 95% CI. on the value of κ. Panels 9F, G, and H show the histograms and 95% CI; notice how the CI's have widened. Because the fit is now on a data set that has not reached saturation, the prediction of the eventual saturation κ is less precise.

When performing a bootstrap analysis in Loglet Lab, the user should inspect the residuals for outliers and other suspect data points, because an anomalously large residual value can unduly undermine the bootstrap. If $\mathbf{D}_0$ is very noisy or contains many outliers, then each D_i^* synthesized from it will be even noisier, producing unrealistic CI's.

6. The Logistic Substitution Model

In our Introduction we promised that loglets could analyze the rise, leveling, and fall of competitors substituting for one another. The "species" do not have to be organisms in an ecosystem; rather, they can be technologies and products competing in a market. For example, we can think of different modes of transportation (such as horses, trains, cars, and airplanes) as competing in the same market. Our discussion will focus on competing technologies rather than species.

The **logistic substitution model** describes the fraction of the niche or market share of the competitors. The life cycle of a competitor can be partitioned into three distinct phases: growth, saturation, and decline. The growth and decline phases represent logistic growth processes, which as we will see, influences the saturation phase.

The assumptions behind the logistic substitution model, as developed by Nakicenovic and Marchetti[19, 20], are:

- New technologies enter the market and grow at logistic rates.
- Only one technology saturates the market at any given time.
- A technology in saturation follows a non-logistic path that connects the period of growth to its subsequent period of decline.
- Declining technologies fade away steadily at logistic rates uninfluenced by competition by new technologies.

The first assumption implies that growth can be modeled with an S-shaped logistic. The fourth also implies that the decline phase can also be modeled with a logistic with a negative Δt. The second and third allow us to determine saturation behavior by competition from emerging technologies.

As an example, we apply the logistic substitution model to the American recording media market. (This example is also featured in the Tutorial; these data are included with the Loglet Lab software, along with other data which illustrate concepts of logistic analysis.) The shift in market dominance from vinyl records (LPs) to cassettes to compact discs (CDs) proved to be a remarkably orderly substitution, as can be seen in Figure 10 which plots the number of LPs, cassettes, and CDs sold from 1977 to 1996. Because the logistic substitution model uses market shares as

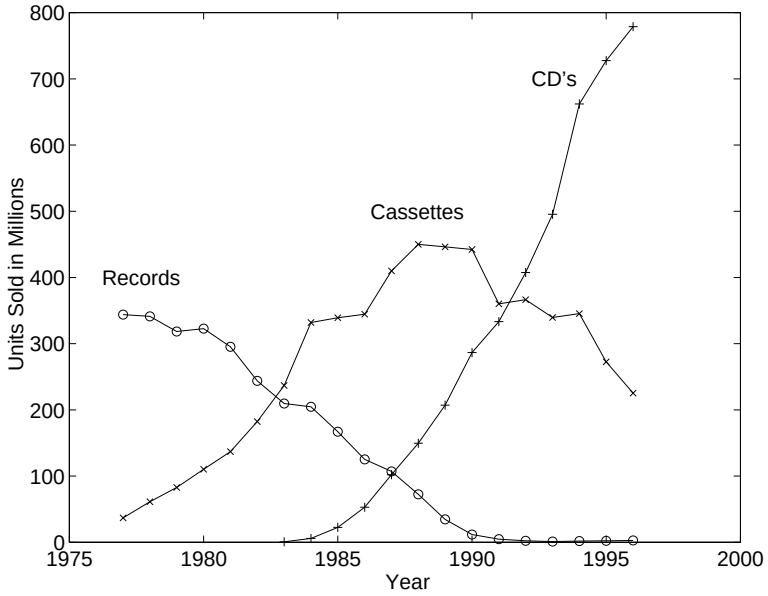

FIGURE 10. Unit sales of U.S. music recording media. Source of data: [21].

opposed to absolute, it is general enough to be useful for illustrating *competition* of species, products, or technologies in systems where the assumptions of logistic growth and decline hold.

6.1. **Numerical methods for logistic substitution.** Now we present the mathematical methods used by the logistic substitution modeling engine.

Suppose we have n technologies competing in a market over the course of m years. Our data representation now becomes a set of n vectors of length m, which can be stored in an $m \times n$ matrix column-wise. To proceed with the model, the first step is to transform the raw data, d_{ij}, into fractional shares $\mathbf{Fr}_{ij}$ of the market:

$$\mathbf{Fr}_{ij} = \frac{d_{ij}}{\sum_{k=1}^{n} d_{kj}}$$

Secondly, the fractional shares $\mathbf{Fr}_{ij}$ are transformed using the previously described Fisher-Pry transform

$$F_{FP\,ij} = FP(\mathbf{Fr}_{ij}) = \frac{\mathbf{Fr}_{ij}}{1 - \mathbf{Fr}_{ij}}.$$

Recall that under the Fisher-Pry transform, an S-shaped logistic becomes a straight line when plotted on a semi-log graph; data which grow (or decline) logistically will appear to grow (or decline) linearly when transformed.

The logistic substitution model generates substitution curves, $L_1, L_2, \ldots, L_n$, that correspond to the fractional market share data $Fr_1, Fr_2, \ldots, Fr_n$. The smooth curves follow the market share through the three substitution phases: logistic growth, non-logistic saturation, and logistic decline.

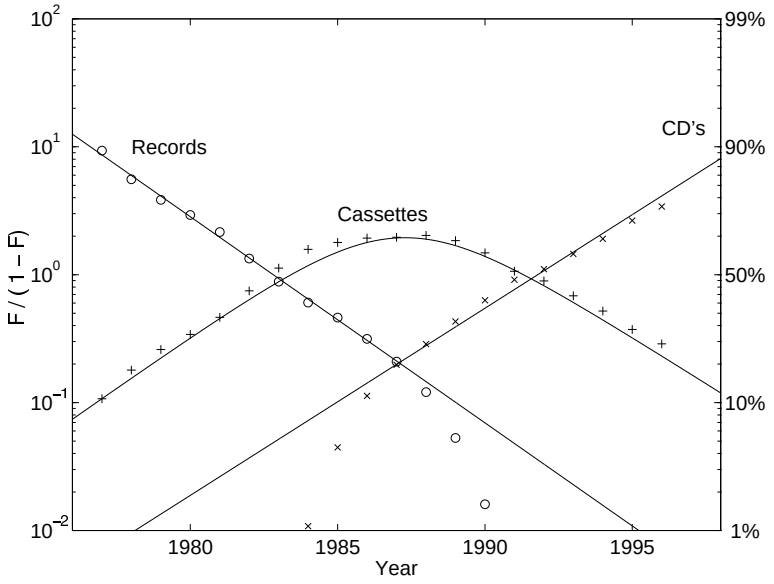

FIGURE 11. Fitted logistic substitution of US music recording media. Source of data: [21].

The first step in generating these curves from the logistic substitution model is to fit a curve to the growth phase of each technology. (Alternatively, as when data for the growth phase is unavailable for a particular technology, we fit a curve to its decline phase.) Reiterating from above, because we are working in the Fisher-Pry transform space, then

$$\ln \frac{L_i}{1 - L_i} = -\frac{\ln 81}{\Delta t_i}(t - t_{mi})$$

is linear, and we can estimate the parameters for such a curve with linear regression. As before, Δt_i is the characteristic growth time for the ith technology, and t_{mi} is the midpoint of the ith technology's period of growth or decline.

Note that for the logistic substitution model, we use a logistic with only two parameters, because the third parameter, saturation level, κ, is fixed at 1, or 100%. Without the introduction of a new technology, the last technology in the growth phase would grow to a 100% market share. If a new technology is introduced, its growth must come at the cost (primarily) of the leading technology, causing it to saturate and decline.

For each technology, it is necessary to specify the time window that represents its logistic growth (or decline). Within this time window are the data that will be used to estimate the parameters for the logistic that we just described. Figure 11 shows the logistic substitution model for the recording media data. The years 1975 to 1985 were used to estimate the logistic decline phase of LPs. The years 1977 to 1985 were used to model the logistic growth phase of cassettes. The years 1988 to 1996 were used to estimate the parameters for the logistic growth phase of CDs.

The growth and decline phases can be represented by logistic curves, but this is not the case for the saturation phase. Because only one technology, L_s, can be

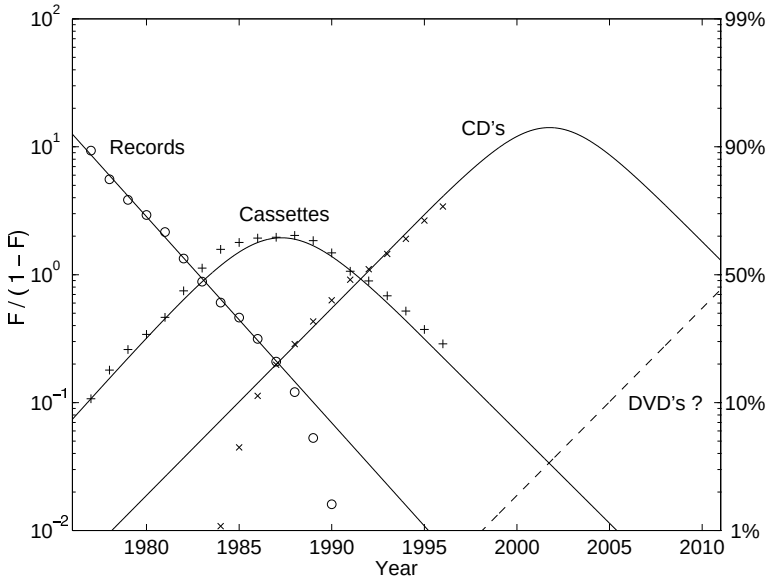

FIGURE 12. Fitted logistic substitution of US music recording media with the introduction of a hypothetical new technology (DVDs). Source of data: [21].

saturating at a time, its market share can be calculated by subtracting the sum of the shares of all the other technologies–which must be known, since they must be either growing or declining–from unity (100%):

$$L_s = 1 - \sum_{i \neq s} L_i.$$

How do we know when each phase begins or ends? If

$$y_s(t) = \ln \frac{L_s(t)}{1 - L_s(t)},$$

then the termination of the saturation phase comes at time t at which

$$\frac{y_s''(t)}{y_s'(t)} \text{is at a minimum.} \tag{12}$$

When the saturation phase for a technology ends, it proceeds directly into its decline phase, and the saturation phase for the next technology immediately commences. The two parameters for the logistic decline phase of the curve are given by:

$$\begin{aligned} \Delta t_s &= \frac{\ln(81)}{y_s'(t)} \\ t_{m_s} &= \ln \frac{(y_s(t) - \frac{ln(81)}{\Delta t}\, t)}{\frac{\ln(81)}{\Delta t}} \end{aligned}$$

In Figure 11 the LPs are in the logistic decline phase, cassettes are in transition, and CDs are in logistic growth.

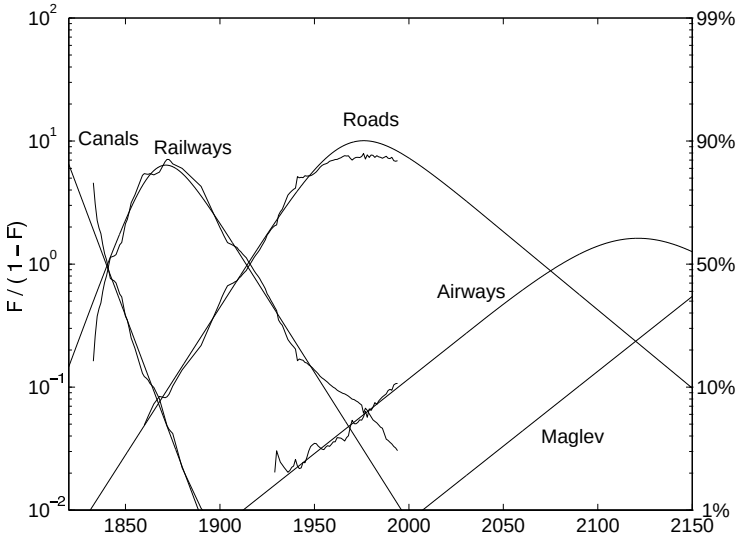

FIGURE 13. Fitted logistic substitution of US transportation infrastructure. After [22].

6.2. **Assessing the possible impact of new technologies.** One motivation in studying history is anticipating the future. We shall now discuss how to visualize the impact of new technologies on market shares.

Rather than using regression, we can specify the parameters for a logistic. This is useful when few or no data are available, as is the case for new technologies seeking to be market leader.

Continuing with the example of recording media, we consider a prospective competitor to follow CDs: the digital versatile disk, or DVD, which has the same size as a CD with about 10 times the storage capacity. We used Loglet Lab to estimate the Δt_i and t_{m_i} that characterize the speed and span of the rise of LPs, cassettes, and CDs. Based on these numbers, we could conjecture similar values of Δt_4 and t_{m_4} for DVDs. We chose a Δt_4 of 13 years and a t_{m_4} of 2012 to generate Figure 12, which shows how the new competitor would affect the market that we had presented in Figure 11.

Figure 13 shows the example of competition and substitution among canals, rails, roads, airways, and a possible competitor, magnetically levitated trains (maglevs)[6]. Here, market share is the length of existing canals, rails, roads, and air routes, not unit sales. This figure also demonstrates familiar competitors (modes of transportation) over a much longer time span than the recording media example and increases the competitors to five. The estimated rates of change of kilometers of each competitor and of the whole American transportation infrastructure illuminate the forces at work. This history of substitution begins in the era of the Erie Canal and extends a century and half into the future.

6.3. **Implementation in Loglet Lab.** The Loglet Lab software condenses the logistic substitution model into two steps. First, it performs a Fisher-Pry transform

[6]For an analysis of US transport infrastructure using the logistic substitution model, see [22].

on all of the data to assist in the identification of the growth (and decline) phases. Second, it asks the user to give either a time window for the growth (or decline) phase or a set of parameters (Δt and t_m) for each technology. Using this input, the logistic substitution engine fits a curve to the growth (or decline) phase of each technology, determines the saturation point based on the criterion in equation (12), and plots the substitution curves.

Loglet Lab can accommodate a large number of data sets, so users can easily add one or more hypothetical competitors and envision several different scenarios for the evolving markets.

7. Conclusion

The Loglet Lab software package provides a means of running the models as described in this paper by making the mechanisms behind the model transparent to the user. Ease of use allows us to envision several scenarios in a manner of minutes, as we can quickly run these models several times with different parameterizations. Combining this with a diverse range of visualizations, Loglet Lab offers a rich analysis of data sets that exhibit complex growth.

Yet underlying the complexity of these models are parameters with concrete and, more importantly, informative interpretations. The emphasis of loglet analysis becomes not building or running the model, but interpreting it.

Acknowledgements: We thank Arnulf Gruebler, John Helm, Cesare Marchetti, Nebojsa Nakicenovic, Paul Waggoner, and Iddo Wernick.

Appendix A. Loglet Lab Software Implementation

The algorithms and methods described in this paper were originally implemented in ANSI C and the mathematics package MatLab [23] in a Unix computer environment. Many of the algorithms are based on the C source code presented in Press, et al. [24] and Engeln-Müllges-Uhlig [25].

The Loglet Lab software project's goal is to build a graphical user interface (GUI) for these algorithms for use on a modern operating system, namely Microsoft Windows 95 and NT (collectively, the "Win32" platform). To this end, Loglet Lab was built in Microsoft's Visual C++ integrated development environment and features a spreadsheet control developed by Visual Components that provides a familiar interface for data entry. Loglet Lab allows users to cut-and-paste their data between spreadsheets such as Excel. Appendix B describes the current state of this project.

Other advanced regression methods (e.g., Marchetti's method for robust weighted regression, or $\mathcal{CM}$ regression) have been implemented in C or MatLab, but they have yet to be ported to the Win32 environment.

Appendix B. Current Status of the Loglet Lab for Windows Software Package

As of 1 July 1999, Loglet Lab for Windows 95 and Windows NT is up to version 1.1.6. A Tutorial for Loglet Lab is also available [10]. The software is available from our web site: `http://phe.rockefeller.edu`. Using this version of Loglet Lab, a user can:

- fit one or more logistics, including declining logistics, to time series data of arbitrary length using a non-linear least-squares regression algorithm as described in section 5. For this purpose, a "Logistic Fitting Wizard" can suggest initial values for the parameters of each logistic as the point of departure for the fitting engine;
- plot the data with its fitted component logistic(s);
- plot the residuals as described in section 4.4;
- plot the data with the Fisher-Pry transform applied, linearizing the time-series data and its fitted logistic(s);
- plot the "bell" view described in section 4.3, which graphs the derivative (rate of change) of a component logistic and the discrete rate of change of the data;
- determine confidence regions for each loglet parameter by using the Bootstrap method, a modern technique which utilizes Monte Carlo methods to determine the statistical error associated with each fitted parameter.
- apply a logistic substitution model to multiple sets of time-series data, including the capacity to add hypothetical competitors.
- extrapolate a model into the past or future;
- annotate the graph (i.e., add a title and label the axes);
- print graphs using any supported Windows printer;
- export or import data to and from Microsoft Excel, or as plain text;
- consult online Help files; and
- peruse through a "Data Gallery" of "classic" sets of logistic time-series data.

Possible improvements for future versions of Loglet Lab include:

- adding an "Undo" function so users can correct their mistakes;
- alternative regression methods (robust, $\mathcal{CM}$);
- export a graph for use on the World Wide Web or in presentation applications such as Microsoft PowerPoint;
- adding the option of constraining parameters to a range of values, as opposed to holding to a single constant value;
- alternative algorithms for providing initial values of parameters;
- visual fitting of component logistic models, where the user interactively changes the parameters using a mouse. A prototype for this can be found at `http://phe.rockefeller.edu/applets/index.html`;
- improving spreadsheet functionality (e.g., formulas); and
- improving the general user interface.

We hope users will send us suggestions to improve Loglet Lab, the Tutorial, and this paper. Please send comments (and bug reports) to `loglet@phe.rockefeller.edu`. We will also post information on the status of Loglet Lab on our web site: `http://phe.rockefeller.edu/`.

References

[1] Marchetti, C., Meyer, P. S., and Ausubel, J. H.: Human Population Dynamics Revisited with the Logistic Model: How Much Can Be Modeled and Predicted? *Technological Forecasting and Social Change*, 52, 1–30 (1996).

[2] Ausubel, J. H. and Marchetti, C.: Elektron: Electrical systems in retrospect and prospect. *Daedalus*, 125(3), 139–169 (1996).

[3] Grübler, A.: *The Rise and Fall of Infrastructures*. Springer–Verlag, New York, 1990.

[4] Nakicenovic, N.: *Growth to Limits: Long Waves and the Dynamics of Technology*. PhD thesis, University of Vienna, Vienna, Austria, 1984.

[5] Marchetti, C.: Society as a Learning System. *Technological Forecasting and Social Change*, 18, 267–282 (1980).

[6] Fisher, J. and Pry, R.: A simple substitution model of technological change. *Technological Forecasting and Social Change*, 3, 75–88 (1971).

[7] Meyer, P. S.: Bi–Logistic Growth. *Technological Forecasting and Social Change*, 47, 89–102 (1994).

[8] Nakicenovic, N. and Grübler, A., eds.: *Diffusion of Technologies and Social Behavior*. Springer–Verlag, Berlin, Germany, 1991.

[9] Kindleberger, C. P.: *World Economic Perspectives: 1500 to 1990*. Oxford University Press, Oxford, 1996.

[10] Yung, J. W., Meyer, P. S., and Ausubel, J. H.: The loglet lab software: A tutorial. *Technological Forecasting and Social Change*, 61(3), 273–295 (1999).

[11] Montroll, E. W.: Social dynamics and the quantifying of social forces. *Proceedings of the National Academy of Sciences (USA)*, 75(10), 4633–4637 (1978).

[12] Lotka, A. J.: *Elements of Physical Biology*. Williams and Wilkins, Baltimore, MD, 1925. (republished as *Elements of Mathemtical Biology*, Dover, New York, 1956).

[13] Kingsland, S. E.: *Modeling Nature: Episodes in the History of Population Ecology*. University of Chicago Press, Chicago, 1985.

[14] Ince, E. L.: *Ordinary Differential Equations*. Dover, New York, 1956.

[15] Banks, R. B.: *Growth and Diffusion Phenomena: Mathematical Frameworks and Applications*. Springer-Verlag, Berlin, Germany, 1994.

[16] Thornton, H. G.: On the development of a standardized agar medium for counting soil bacteria. *Ann. Appl. Biol.*, 9, 241–274 (1922).

[17] Wilson, E. B.: The logistic or autocatalytic grid. *Proceedings of the National Academy of Sciences USA*, 11(8), 451–457 (1925).

[18] Efron, B. and Tibshirani, R. J.: *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY, 1993.

[19] Marchetti, C. and Nakicenovic, N.: The dynamics of energy systems and the logistic substitution model. *IIASA Research Report RR–79–13*, (1979). International Institute for Applied Systems Analysis, Laxenburg, Austria.

[20] Nakicenovic, N.: Software package for the logistic substitution model. *IIASA Research Report RR–79–12*, (1979). International Institute for Applied Systems Analysis, Laxenburg, Ausutria.

[21] RIAA.: Annual report. Technical report, Recording Industry Association of America, Washington, D.C., 1997.

[22] Ausubel, J. H., Marchetti, C., and Meyer, P. S.: Toward green mobility: The evolution of transport. *European Review*, 6(2), 143–162 (1998).

[23] Moler, C.: *Matlab 4.0 Reference Guide*. The Mathworks Inc., Natick, MA, 1996. (`http://www.mathworks.com`).

[24] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P.: *Numerical Recipes in C: The Art of Scientific Computing*. 2 ed., Cambridge University Press, 1992.

[25] Engeln-Müllges, G. and Uhlig, F.: *Numerical Algorithms with C*. Springer–Verlag, New York, 1996.

Received 5 February 1999