1 Introduction to simulation of biological systems

Overview

This chapter is built around analyzing a real data set obtained from a real biological system to illustrate several complementary approaches to simulation and analysis. The particular system studied (a home aquarium) is a well-mixed chemical reactor. Or, more accurately, the system studied is treated as a well-mixed chemical reactor, a basic modeling paradigm that will appear again and again in this book.

Here, we look at this single physical system from several different perspectives (that is, under different sets of underlying modeling assumptions) with the aim of motivating the reader to undertake the study of the rest of this book. The aim is not to overwhelm the reader with mathematical details that can be found in later chapters. Therefore let us clearly state at the outset: it is not expected or required that the reader follow every detail of the examples illustrated here. Instead, we invite the reader to focus on the basic assumptions underlying the methods applied, and to compare and contrast the results that are obtained based on these different approaches. Proceeding this way, it is hoped that the reader may gain an appreciation of the breath of the field. Furthermore, it is hoped that this appreciation will continue to grow with a study of the rest of this book and beyond.

1.1 Modeling approaches

The number of different approaches to simulating biosystems behavior is perhaps greater than the number of biological systems. The number is at least large enough that a finite and complete list cannot be constructed. Simulation methods may be classified according to the physical systems simulated (for example, cellular metabolism, whole-body drug distribution, or ecological network dynamics), the sets of assumptions used to build a simulation (for example, rapid mixing versus spatial inhomogeneity in chemical reaction systems), or the mathematical/computational formulation of the simulation (for example, systems of ordinary differential equations versus statistical inference networks for describing

regulation of gene transcription). A glance at the table of contents reveals that most of this book is organized by biological system or application area. (Modeling assumptions and relevant computational techniques are introduced as necessary.)

These biological systems can (and will!) be studied by applying a variety of sets of assumptions and associated computational methods. Doing this, we will see that the methodology applied to a given system depends strictly on what one thinks one knows about the system in advance, and what one wishes to discover through computational analysis. In the following introductory example we will see that what we can learn (for example, what variables and what parameters we can estimate) depends on the prior knowledge built into a model, including (but not limited to) what data are available for a given system.

1.2 An introductory example: biochemistry of a home aquarium

As our first exemplar modeling study, let us analyze the buildup and reaction of waste materials in a home aquarium, a system that may be familiar to some readers. Ammonia (NH_3), which is toxic to fish, is excreted from fish as a waste product and produced through decomposition of organic matter. In a well-functioning aquarium, nitrifying bacteria in the aquarium filter oxidize ammonia to nitrite (NO_2^-) and oxidize the nitrite to nitrate (NO_3^-). Of these three nitrogen-containing compounds, nitrate is by far the least toxic to fish.

When one sets up a new aquarium, populations of nitrifying bacteria are yet to be established, and concentrations of toxic compounds can temporarily build up. Figure 1.1 plots data collected by the author from his own aquarium following the addition of fish into a previously uninhabited new tank. Here we see that ammonia concentration tends to build up over the first week or more. Once significant populations of bacteria that convert NH_3 to NO_2^- appear, the ammonia declines while the nitrite level increases. Nitrite concentration eventually declines as nitrate begins to appear.[1]

We wish to understand how these three concentrations are related kinetically. To simplify the notation, we introduce the definitions $x_1 =$ [ammonia], $x_2 =$ [nitrite], $x_3 =$ [nitrate] for the concentration variables. As already described, the expected sequence of reaction in this system is $x_1 \rightarrow x_2 \rightarrow x_3$. In fact, that sequence is apparent from the data illustrated in the figure. Ammonia (x_1) peaks around day 10, followed by nitrite (x_2) around day 13. Nitrate concentration (x_3) really picks up following the peak in nitrite, and continues to steadily increase.

[1] Ammonia, nitrite, and nitrate exist in aqueous solution in a number of rapidly interconverting forms. For example, at low pH NH_3 is largely protonated to form the ammonium ion NH_4^+. Here, the terms ammonia, nitrite, and nitrate are understood to include all such rapidly converting species of these reactants.

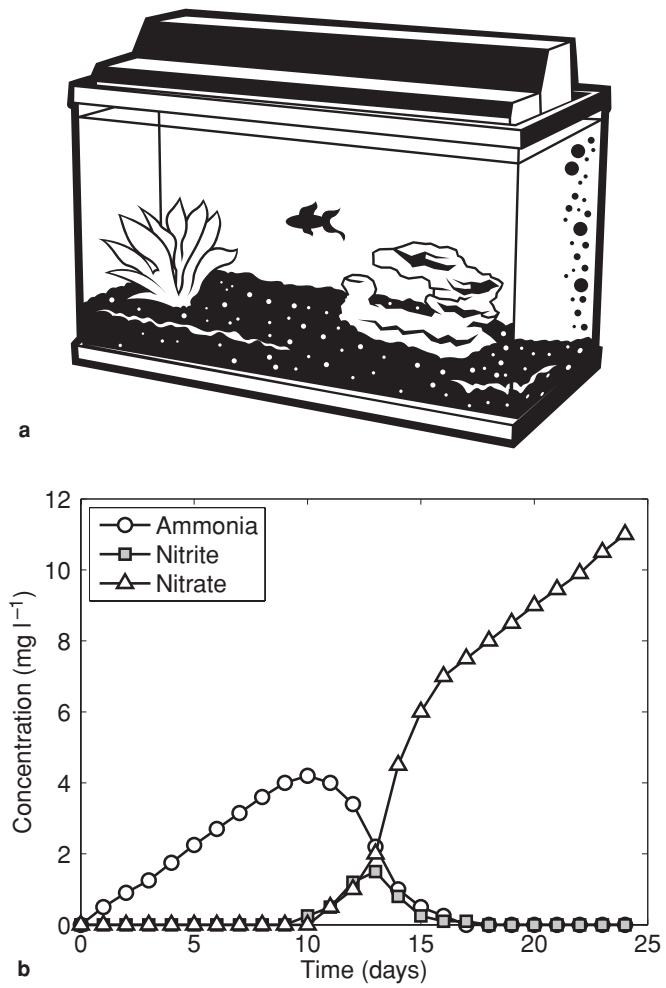

Figure 1.1

A home aquarium. The plot in panel (b) shows ammonia, nitrite, and nitrate concentration versus time in a home aquarium. Concentrations are given in units of mg of nitrogen per liter (mg l^{-1}).

Yet in addition to the reaction sequence, is it possible to obtain additional quantitative information from these data? To do so, let us construct a series of simple models and see what we can find.

1.2.1 First model: a nonmechanistic analysis

In the first model we would like to introduce the minimum number of assumptions that allow us to explain the observed data. The idea is to construct a general set of

governing differential equations for x_1, x_2, and x_3 based only on the assumption of mass conservation. Conservation laws are a universal tool for developing models of many systems, including practically all of the examples studied in this book. Here we assume that the volume of the system, V, remains constant. So the rate of change of mass of a substance, $\frac{d(cV)}{dt}$, is equal to $V\frac{dc}{dt}$. In this case, the general differential equation for concentration of a reactant is

$$V\frac{dc}{dt} = \begin{pmatrix}\text{rate of production or}\\ \text{input measured in units}\\ \text{of mass per unit time}\end{pmatrix} - \begin{pmatrix}\text{rate of loss}\\ \text{measured in units of}\\ \text{mass per unit time}\end{pmatrix} \tag{1.1}$$

or

$$\frac{dc}{dt} = \begin{pmatrix}\text{rate of production or}\\ \text{input measured in units}\\ \text{of mass per unit time}\\ \text{per unit volume}\end{pmatrix} - \begin{pmatrix}\text{rate of loss}\\ \text{measured in units of}\\ \text{mass per unit time}\\ \text{per unit volume}\end{pmatrix}. \tag{1.2}$$

Applying this general form to the variables x_1, x_2, and x_3 gives

$$\begin{aligned}\frac{dx_1}{dt} &= k(t) - r_{12}(t)\\ \frac{dx_2}{dt} &= r_{12}(t) - r_{23}(t)\\ \frac{dx_3}{dt} &= r_{23}(t),\end{aligned} \tag{1.3}$$

where $k(t)$ is the rate of ammonia (x_1) production, and $r_{12}(t)$ and $r_{23}(t)$ are the rates of conversion from ammonia to nitrite and from nitrite to nitrate, respectively. The first equation states that the rate of change of x_1 is equal to the rate of production minus the rate of degradation. Similar statements of mass conservation follow for dx_2/dt and dx_3/dt. Since no processes degrading nitrate are considered, there is no degradation term in the dx_3/dt equation. Because $k(t)$, $r_{12}(t)$, and $r_{23}(t)$ are (so far) assumed to be arbitrary functions, we have (so far) not introduced any assumptions about the rules governing the behavior of these functions. The names and definitions of the model variables are listed in the table below.

Variable	Units	Description
x_1	mg l^{-1}	concentration of ammonia
x_2	mg l^{-1}	concentration of nitrite
x_3	mg l^{-1}	concentration of nitrate
k	$\text{mg l}^{-1}\,\text{day}^{-1}$	rate of ammonia production
r_{12}	$\text{mg l}^{-1}\,\text{day}^{-1}$	rate of nitrite production from ammonia
r_{23}	$\text{mg l}^{-1}\,\text{day}^{-1}$	rate of nitrate production from nitrite

We call this model "nonmechanistic" because it does not invoke any biochemical/biophysical mechanisms to describe the rates of conversion $r_{12}(t)$ and $r_{23}(t)$, or the rate of production $k(t)$. Instead these rates are all allowed to be arbitrary functions.

So what can we do with this simple general model? One useful thing we can do is analyze the data using the model to estimate $r_{12}(t)$, $r_{23}(t)$, and k and test the model assumptions. From Eq. (1.3), we have $r_{23}(t) = dx_3/dt$, which can be numerically approximated using a finite difference approximation

$$\hat{r}_{23}(t) \approx \frac{x_3(t + \Delta t) - x_3(t - \Delta t)}{2\Delta t}. \tag{1.4}$$

Here Δt is the discrete time step over which the data in Figure 1.1 are sampled. (I took one measurement per day, so Δt = 1 day.) Equation (1.4) is the "central-difference" approximation for the derivative of x_3 with respect to time.[2] Here we use the notation $\hat{r}_{23}$ to denote the approximation (from the data) of r_{23}. Next, given our approximation of $r_{23}(t)$, we can approximate $r_{12}(t)$:

$$\hat{r}_{12}(t) = \frac{dx_2}{dt} + r_{23}(t) \approx \frac{x_2(t + \Delta t) - x_2(t - \Delta t)}{2\Delta t} + \hat{r}_{23}(t). \tag{1.5}$$

Similarly, we can approximate k as a function of time

$$\hat{k}(t) = \frac{dx_1}{dt} + r_{12}(t) \approx \frac{x_1(t + \Delta t) - x_1(t - \Delta t)}{2\Delta t} + \hat{r}_{12}(t). \tag{1.6}$$

Values of $\hat{k}(t)$, $\hat{r}_{12}(t)$, and $\hat{r}_{23}(t)$ computed from the data in Figure 1.1 are plotted in Figure 1.2.[3] From these estimated rates we learn a number of things about this system not immediately apparent from a simple inspection of the raw data. First, we can see that the rate of ammonia production ($\hat{k}(t)$) is estimated to be approximately constant. Moreover, this analysis provides an estimate of the constant k, approximately 0.4 to 0.5 mg l^{-1} day^{-1}. This observation is perhaps not unexpected, because the number and size of the fish remained approximately constant over the course of the experiment, as did the amount of food introduced per day. Therefore we might have expected the rate of ammonia production to be nearly constant. Second, the analysis reveals that nitrite production ($\hat{r}_{12}(t)$) peaks near day 13 while nitrate production peaks shortly after, around day 14. Towards the end of the experiment, all of the reaction rates converge to equal approximately 0.5 mg l^{-1} day^{-1}. Finally, we note that the estimated rates $\hat{k}(t)$, $\hat{r}_{12}(t)$, and $\hat{r}_{23}(t)$ remain positive for the duration of the experiment. This observation makes sense, because under normal conditions neither of the nitrification reactions is expected

[2] Discrete approximations of derivatives are reviewed in Section 9.1 in the Appendices.
[3] Computer codes (implemented in MATLAB) for this and all of the examples in this book can be found online at the URL http://www.cambridge.org/biosim.

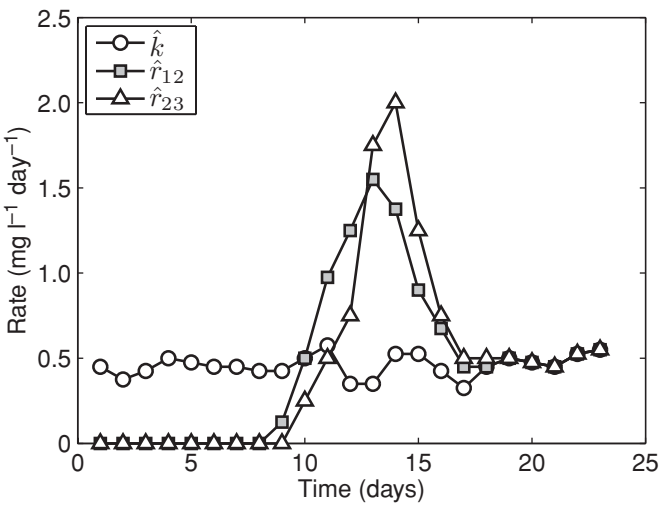

Figure 1.2

Values of $\hat{k}(t)$, $\hat{r}_{12}(t)$, and $\hat{r}_{23}(t)$ estimated from Eqs (1.4)–(1.6) and the data in Figure 1.1. These rates are expressed in units of mass of nitrogen per unit volume per unit time: mg l^{-1} day^{-1}.

to proceed in the reverse direction. Thus the result that $\hat{r}_{12}(t)$ and $\hat{r}_{23}(t)$ remain positive provides a useful check of the physical realism of the model.

To summarize, analyzing the data of Figure 1.1 using the simple model of Eq. (1.3), which invokes no more serious assumption than conservation of mass, provides quantitative estimates of a number of variables that are not directly measured.

1.2.2 Nonmechanistic analysis with noise

The preceding analysis was applied to a relatively noise-free data set, yielding reasonable (and smooth) numerical estimates for the derivatives in Eqs (1.4)–(1.6). However, differentiation has the unfortunate side effect of tending to magnify noise. And since significant measurement noise is often associated with real-world biological signals, analyses that require the estimation of derivatives of biological data are often seriously confounded by noise.[4]

[4] The aquarium experiment studied here cannot be regarded as a precisely controlled study. The original data set of Figure 1.1 was collected by the author using a simple consumer kit, with which the concentrations are estimated by visual comparison of the fluid in a test-tube assay with a color chart, possibly introducing bias. Although dilutions and replicates were performed as appropriate, there is a human psychological component to interpreting these assays. Furthermore, because the nitrate assay used is relatively insensitive over the reported concentration range, data were obtained by a combination of interpolation, assuming a constant total nitrate production rate, and colorimetric assay. Given the potential for bias, the reader is encouraged to conduct his or her own experiments in his or her own home laboratory!

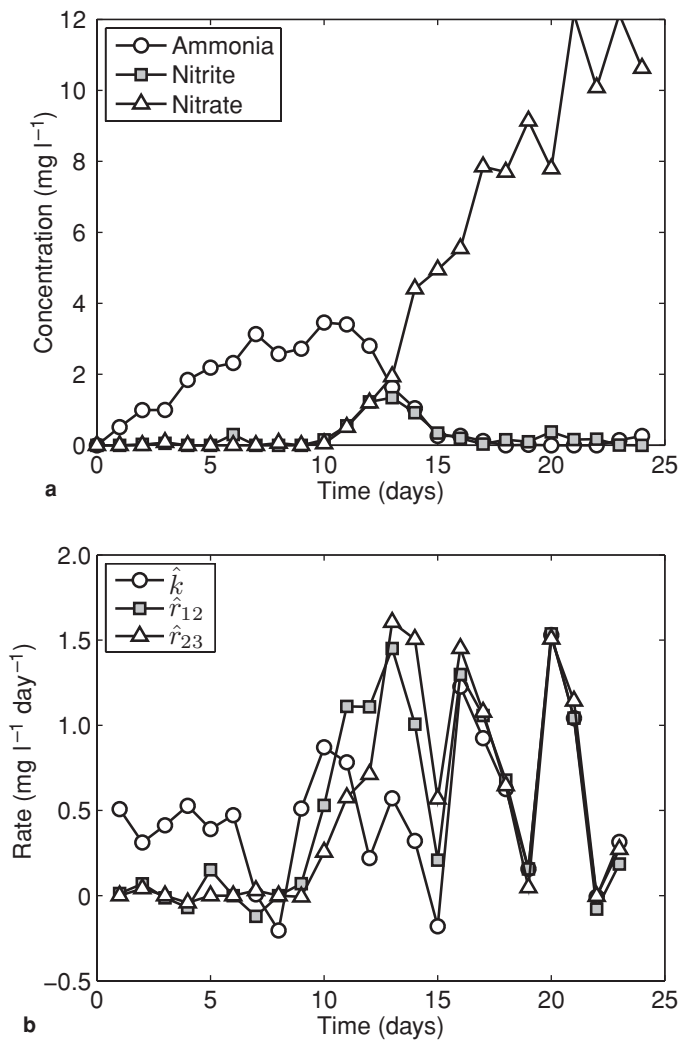

Figure 1.3

Analysis of noisy aquarium data. In panel (a) the data of Figure 1.1 are reproduced with added noise. The analysis of the previous section is reproduced in panel (b), with $\hat{k}(t)$, $\hat{r}_{12}(t)$, and $\hat{r}_{23}(t)$ computed from Eqs (1.4)–(1.6) applied to the noisy data from panel (a).

To illustrate this problem, and to explore some ideas of how to deal with it, we can add some noise to our aquarium data. Figure 1.3(a) shows the same data as those of Figure 1.1, with a relatively small amount of noise added. We can see that the basic trends in the data remain the same, but this data set is less smooth than the previous one.

The values of $\hat{k}(t)$, $\hat{r}_{12}(t)$, and $\hat{r}_{23}(t)$, computed from Eqs (1.4)–(1.6) for these data are plotted in Figure 1.3(b). Here we can see the consequence of differentiating

the noisy signals from the upper panel: these estimated rates are terribly noisy, and hardly resemble the estimates illustrated in Figure 1.2. This analysis tells us very little about what we think happened in the experiment. With only a modest amount of random noise added to the signal, we are no longer able to draw any conclusions or confidently estimate any of the unmeasured variables in the system.

To analyze the noisy data effectively requires additional analysis. One place to start is to reexamine the data to look for clues on how to improve our calculations. Is there any trend in the data set of Figure 1.3 that we might be able to take advantage of? After a careful look, the observer's attention might be directed to the fact that over the second half of the experiment the ammonia and nitrite concentrations remain very small compared with nitrate, which continues to grow. If x_1 and x_2 remain constant over some time regime (say in the limit $t \to \infty$ or, more practically, for the last two weeks of the experiment), then in this time window Eq. (1.3) reduces to

$$\begin{aligned} \frac{dx_1}{dt} &= k(t) - r_{12}(t) = 0 \\ \frac{dx_2}{dt} &= r_{12}(t) - r_{23}(t) = 0 \\ \frac{dx_3}{dt} &= r_{23}(t). \end{aligned} \tag{1.7}$$

This system of equations tells us that $dx_3/dt = k(t)$ in this time window. Furthermore, observation of the raw data in Figure 1.3 tells us that the rate of growth of nitrate (or dx_3/dt) is approximately constant in this time window.

Therefore, before trying to estimate the rates $k(t)$, $r_{12}(t)$, and $r_{23}(t)$, there is justification for introducing the a priori assumption that $k(t)$ becomes constant *at some point in the experiment*. If we have other reasons (such as those discussed above) to think that $k(t)$ might be constant *throughout the whole experiment*, we might hypothesize that this is the case. Doing so, this hypothesis can be formally built into the analysis as an additional assumption, while being sure to remember that this assumption is a hypothesis that remains to be tested against the data.

To test the hypothesis, we can sum the equations in Eq. (1.3) to obtain

$$\frac{d}{dt}(x_1 + x_2 + x_3) = k(t). \tag{1.8}$$

If indeed $k(t)$ is constant then the sum $x_1 + x_2 + x_3$ should increase at at constant rate throughout the experiment. In Figure 1.4 we plot this sum to test the hypothesis and find that, indeed, $x_1 + x_2 + x_3$ increases at an approximately constant rate. The solid line in the figure represents a line of slope $\hat{k} = 0.45$ mg l^{-1} day^{-1}, which is the estimate of the constant rate of ammonia production obtained from this analysis. With the hypothesis that $k(t)$ is constant not disproved, and an estimate

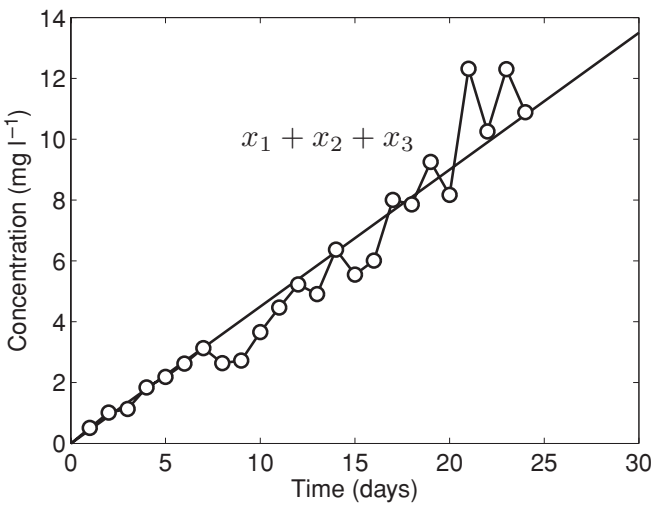

Figure 1.4

Plot of summed data $x_1 + x_2 + x_3$. The solid line has slope 0.45 mg l^{-1} day^{-1}, which provides an estimate of the constant k.

of k in hand, we next continue by using this assumption to help estimate the other rates.

With constant k, Eq. (1.3) provides three equations for the two unknowns $r_{12}(t)$ and $r_{23}(t)$:

$$\begin{aligned} r_{12}(t) &= k - \frac{dx_1}{dt} \\ r_{12}(t) - r_{23}(t) &= \frac{dx_2}{dt} \\ r_{23}(t) &= \frac{dx_3}{dt} \end{aligned} \tag{1.9}$$

with the equivalent numerical approximation

$$\begin{aligned} \hat{r}_{12}(t) &= \hat{k} - \frac{x_1(t+\Delta t) - x_1(t-\Delta t)}{2\Delta t} \\ \hat{r}_{12}(t) - \hat{r}_{23}(t) &= \frac{x_2(t+\Delta t) - x_2(t-\Delta t)}{2\Delta t} \\ \hat{r}_{23}(t) &= \frac{x_3(t+\Delta t) - x_3(t-\Delta t)}{2\Delta t}. \end{aligned} \tag{1.10}$$

In general this is an ill-posed problem, and there is no solution (for $\hat{r}_{12}(t)$ and $\hat{r}_{23}(t)$) that satisfies all of the equations. Instead, we can seek a solution that in some way approximately solves Eq. (1.10).

Putting this linear system into matrix-vector form, we have

$$\begin{bmatrix} 1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{r}_{12}(t) \\ \hat{r}_{23}(t) \end{bmatrix} = \begin{bmatrix} \hat{k} - \dfrac{x_1(t+\Delta t) - x_1(t-\Delta t)}{2\Delta t} \\ \dfrac{x_2(t+\Delta t) - x_2(t-\Delta t)}{2\Delta t} \\ \dfrac{x_3(t+\Delta t) - x_3(t-\Delta t)}{2\Delta t} \end{bmatrix}. \tag{1.11}$$

One approach to computing $\hat{r}_{12}(t)$ and $\hat{r}_{23}(t)$ is to find the solution that minimizes the error between the left-hand and right-hand sides of this equation.

In fact, there exists a handy general solution to problems of this sort when the error is formulated as the sum of squares of differences. These problems are called *least-squares* problems in mathematics, and here we consider the specific problem of minimizing the error in the overdetermined linear system

$$\mathbb{A}\mathbf{x} = \mathbf{b},$$

where $\mathbb{A}$ is a matrix in which the number of rows (number of equations) outnumbers the number of columns (number of unknowns). Note that the matrix in Eq. (1.11) is a matrix of this type. The least-squares solution (the vector $\mathbf{x}$ that minimizes $\|\mathbb{A}\mathbf{x} - \mathbf{b}\|^2$) is found as the vector $\mathbf{x}$ that solves the well-posed problem

$$\mathbb{A}^T\mathbb{A}\mathbf{x} = \mathbb{A}^T\mathbf{b}.$$

(This least-squares analysis is reviewed in Section 9.2 of the Appendices.)

Applying this formula to Eq. (1.11), we obtain the following estimates for $\hat{r}_{12}(t)$ and $\hat{r}_{23}(t)$.

$$\begin{aligned} \hat{r}_{12}(t) &= \frac{1}{3}\left(2\hat{k} - 2\frac{x_1(t+\Delta t) - x_1(t-\Delta t)}{2\Delta t} + \frac{x_2(t+\Delta t) - x_2(t-\Delta t)}{2\Delta t}\right. \\ &\qquad \left. + \frac{x_3(t+\Delta t) - x_3(t-\Delta t)}{2\Delta t}\right) \\ \hat{r}_{23}(t) &= \left(\hat{r}_{12}(t) + \frac{x_3(t+\Delta t) - x_3(t-\Delta t)}{2\Delta t} - \frac{x_2(t+\Delta t) - x_2(t-\Delta t)}{2\Delta t}\right) \Big/ 2. \end{aligned} \tag{1.12}$$

(This solution is easy to verify and is the subject of Exercise 1.1.)

Holding $\hat{k}$ constant and computing $\hat{r}_{12}(t)$ and $\hat{r}_{23}(t)$ at each time point from Eq. (1.12), we obtain the estimates plotted in Figure 1.5. This result is a clear improvement over Figure 1.3. Here we are able to capture the peak production times near 13 days for nitrite and nitrate. However, the estimates are still noisier than those obtained for the low-noise case illustrated in Figure 1.2. In addition, in