Modeling stochasticity in biological data

J.J.M. Bedaux, dept Theoretical Biology, Vrije Universiteit, Amsterdam 2001-05-07

In the biological sciences mathematical models are indispensable. Nowadays much effort is done in building realistic models to describe complex systems in relatively simple formulas. Generally, the resulting formulas are differential equations based on ideas about the mechanisms underlying the observed phenomena. Most models used in practice, are completely deterministic. The DEB model is an example of such an approach. It is deterministic, apart from the stochastic component in death by aging.

When models are confronted with experimental data to test their adequacy or to estimate parameters, there will always be some discrepancy between model and data. Usually little attention is given to the origin of this discrepancy. It is just called measurement error or noise, suggesting that, in case of a 'true' model, the discrepancy could be diminished or even removed by better experimentation. This vision is seldom correct. In most cases there is some inherent stochasticity one can never get rid of. Stochastic data ask for stochastic models, so the deterministic model must be changed or extended to a stochastic model. There are several options to incorporate stochastic elements in deterministic models.

To be more specific, let us look at the following example. Suppose we have a mathematical model describing the dynamics of a pollutant in an animal. The model can be formulated as a differential equation for the internal concentration *C(t)*:

$$C'(t) = k_u c - k_e C(t)$$

where k_u is the uptake rate, k_e the elimination rate and c the (constant) environmental concentration. If we want to estimate parameters from data (t_1,C_1), (t_2,C_2), …, (t_n,C_n), we have to make the model stochastic. The usual regression approach is to add a stochastic variable to the model function $C(t)$, based on the idea of ' additive noise' . See figure 1a. The deviations between observations C_i and model function $C(t_i)$ are considered as stochastic variables $\underline{e}_i$

$$\underline{C}_i = C(t_i) + \underline{e}_i \qquad \text{with } \underline{e}_i \sim N(0, \sigma^2)$$

or

$$\underline{C}_i \sim N(C(t_i), \sigma^2)$$

A lot of methods are available to estimate parameters.

The approach mentioned above are based on a set of fixed parameter values, i.e. they all use one model function for the whole data set. In other words, every data point is supposed to obey the same model function. This is not always realistic. For instance, in case of one observation per animal this is a very questionable assumption. Animals are not equal, due to genetic or site-specific

differences for instance. A more realistic approach is then to give each animal its own set of parameter values. A random sample of animals results in a random sample of parameter-value sets and so in a random sample of model functions. In other words, the parameters are considered to be stochastic. See figure 1b. A third possibility to implement stochasticity is to consider the animal as a deterministic system with a stochastically changing input. Especially if data are from one individual this can be realistic. A fourth possibility is to describe the system in terms of stochastic differential equations. In the case of DEB this could easily lead to inconsistencies because of the conservation laws of mass and energy.

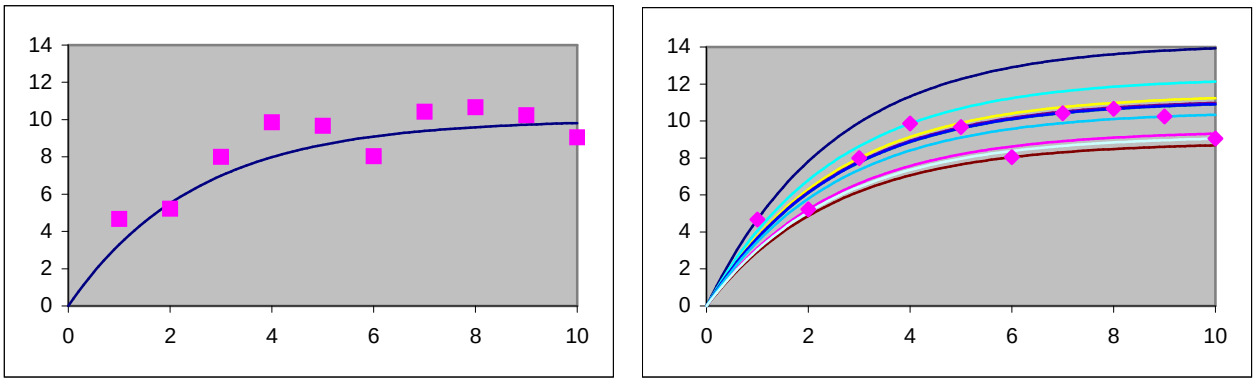

Figure 1. Relation between a data set and the model $C'(t) = k_u c - k_e C(t)$ in two different ways.

a. (left) Regression model. Stochastic deviations between model function and data.

b. (right) Stochastic parameters. Each data point has its own model function.

In this essay we discuss the second approach. In the context of DEB theory this may easily lead to an explosions of the number of parameters. If we look for instance to the description of growth under constant food density (equation 3.20 in the DEB book), we have to deal with the following parameters: V_m, V_h, k_M, v, V_b and X_K. If we want to describe all parameters as stochastic variables, we have to choose classes of distributions. (I think that DEB theory does not give any clue to what kind of distribution is reasonable for this purpose.) Gamma distributions, or log-normal distributions are obvious candidates (they give positive valued real numbers). But then the minimum number of parameters equals 12, if we assume that all these random variables are independent. Actually we deal with some multivariate distribution.

But we have to keep life simple. Not only to keep things manageable, but also to have the complexity of the stochastic part in balance with the complexity of the deterministic part of the model. An escape from this problem is to use the argumentation given on page 334 in the chapter 'Living together'. There variation between parameters of individual organisms of one species is assumed to behave like variation between parameters of different species. That elegant assumption only leads to two additional parameters, if we model the zoom factor z (see Table 8.1) with a log-

normal or a Gamma distribution. Does that give rise to the type of randomness that we observe in experimental data? Therefore we have to work out how the above parameters depend on the zoom factor z. It is easy to see from the definitions that V_h, k_M and v do not depend on z and V_m, X_K and V_b are proportional with z.

What about V_∞? If we consider two individuals with values $V_{\infty 1}$ and $V_{\infty 2}$ the relationship is more complex:

$$V_{\infty 2}^{1/3} = f_2 V_{m2}^{1/3} - V_{h2}^{1/3} = (1 + X_{K1} z/X)^{-1} V_{m1}^{1/3} z - V_{h1}^{1/3}.$$

This cannot be expressed easily in terms of $V_{\infty 2}$. Even if we take food *ad libitum* ($X >> X_K$) we still do not get $V_{\infty 1} = z V_{\infty 2}$, unless we observe ectotherms ($V_h = 0$).

The von Bertalanffy growth rate depends on four of the above parameters, in a non-linear way. That leads to

$$r_{B2} = (3/k_{m1} + 3(1 + X_{K1} z/X)^{-1} V_{m1}^{1/3} z/v_1)^{-1}.$$

If food density is high we get

$$r_{B2} = (3/k_{m1} + 3 V_{m1}^{1/3} z/v_1)^{-1}.$$

The theoretical distribution of (volumetric) length measurements will be very difficult to derive. A simulation study could give more insight into this point, but that is beyond the scope of this essay.

What do we learn from this exercise? DEB theory gives a clue for including stochasticity in deterministic models. The formulae can be worked out, but they will be very difficult to apply. We still have to choose classes of distributions, and in case of more observation per individual, we have to include intra-individual stochasticity to complete the story.