Modeling uncertainty in population biology: how the model is written does matter

Janos G. Hajagos*

March 2, 2004

1 An alternative to second-order Monte Carlo

Monte Carlo based approaches are used to calculate the risk of extinction for threatened species. In the risk assessment the exact values of the statistical moments of the input distributions need to be known. At best, the mean and variance for the growth rate of the population might be known plus or minus ten percent of the estimated value. The normal course of action is to perform a second-order Monte Carlo analysis. In such an analysis, a second statistical distribution is sampled for the moments of the first distribution. Second-order Monte Carlo adds an additional factor of computation time and makes more assumptions about the distribution of moments; when data is sparse, like in the case of endangered species, these additional probabilistic assumptions might not be supported.

An alternative to second-order Monte Carlo analysis is presented in this paper. Instead of sampling from a second statistical distribution, the uncertainty around the moments will be bound, and then propagated through a numerical simulation of population dynamics using interval analysis. With interval analysis no additional assumptions except that the moments are bounded need to be made. It will be shown that there are two ways to write the equation for population growth. The correct equation to use will depend on what is meant by an interval. If one believes that an interval represents a bounded set of possible values then Equation 8 should be used, but if one believes that an interval represents uncertainty of not knowing a fixed value then Equation 9 should be used. The choice is not without consequences: the bounds on the quasi-extinction decline risk will be tighter with Equation 9.

*Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, NY 11794-5245, U.S.A.; email: jhajagos@life.bio.sunysb.edu; fax: 631-632-7626

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

2 Population models

The basic model for growth of an animal population is the exponential growth function, written here in its continuous form

$$f(t) = N_0 \exp(rt), \tag{1}$$

where N_0 is the initial population size, t is time, and r is the per capita rate of growth. This function arises from a solution to the simple differential equation

$$\frac{dN}{dt} = rN, \tag{2}$$

where N is the population size.

Discrete deterministic population models are normally written in the form

$$N_{t+1} = RN_t, \tag{3}$$

where R is a per unit time multiplier, N_t is the population at time t. For predicting N_T, such that, $T \in \{0, 1, 2, \ldots\}$, one has

$$N_{t+T} = R^T N_t. \tag{4}$$

An important relationship exists between R and r, the finite rate of increase and the per capita rate of growth, that is,

$$R = \exp(r). \tag{5}$$

From this point the notation used to write a discrete function of population growth will change. We will now consider the population abundance at time T to be a function of the size of the population at time 0, the time horizon T (the length of the simulation), and the per-capita growth rate r. The equation of population growth rewritten in terms of the new notation is

$$f(N_0, r, T) = N_0 \exp(rT) = N_T. \tag{6}$$

3 Adding stochasticity

For real biological populations, that is, those that are observed in nature, the per-capita rate of population growth is not fixed through time but varies. Equation 6 can be rewritten to take into account varying rates of r

$$f(N_0, \{r_1, \ldots, r_T\}, T) = N_0 \exp\left(\sum_{i=1}^{T} r_i\right) = N_T, \tag{7}$$

where r_i is a random variate from G, a statistical distribution. It is assumed here that G is a normal distribution with a mean $\bar{r}$ and with a standard deviation of σ_r; $r_i = g(\bar{r}, \sigma_r)$ is a random variate from the normal distribution $G(\bar{r}, \sigma_r)$ [Lewontin and Cohen, 1969].

To simulate the potential dynamics to the populations, we make K runs or realizations of the model.

4 Adding measurement uncertainty

To propagate epistemic uncertainty, that is, uncertainty which can reduced through effort, interval analysis [Moore, 1966] will be used. An interval $\mathbf{X}$ is defined as a closed set on the real line, such that, $x \in \mathbf{X} \subseteq \mathbb{R}$ where $\underline{X} \leq x \leq \overline{X}$, and $\underline{X}$ and $\overline{X}$ are the infinimum and supremum, respectively of $\mathbf{X}$. The set of all intervals on the real line is denoted $\mathbb{IR}$. Given intervals $\mathbf{X}$ and $\mathbf{Y}$ addition is defined as

$$\mathbf{X} + \mathbf{Y} = [\underline{X} + \underline{Y}, \overline{X} + \overline{Y}] = \{x + y : x \in \mathbf{X}, y \in \mathbf{Y}\}$$

There are interval definitions for a wide range of basic mathematical operators, such as, $\{-, \times, /, ^2\}$, and for functions, such as, $\{\exp, \log, \sin, \cos\}$. To propagate epistemic errors through a simulation of population dynamics two additional operators need to be defined:

$$\mathbf{X} \times \mathbf{Y} = [\min(\underline{XY}, \overline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}), \max(\underline{XY}, \overline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y})] = \{xy : x \in \mathbf{X}, y \in \mathbf{Y}\}$$

$$\exp(\mathbf{X}) = [\exp(\underline{X}), \exp(\overline{X})] = \{\exp(x) : x \in \mathbf{X}\}.$$

By outwardly rounding the endpoints of an interval operation the interval is guaranteed to contain the true value. For the simulation of population dynamics the Intlab toolbox [Rump, 1999b, Rump, 1999a] for Matlab is used.

The algebra on intervals differs from the algebra on real numbers. For example,

$$\mathbf{C} \times (\mathbf{A} + \mathbf{B}) \subseteq \mathbf{C} \times \mathbf{A} + \mathbf{C} \times \mathbf{B}$$

this is known as the subdistributive law [Moore, 1979]. In the non-strict inequality, equality will hold when $\underline{A}, \underline{B} > 0$. Of more importance is Moore's single use theorem which states that if each variable in a mathematical expression occurs only once then the resulting bounds from applying interval operators will be optimal [Hansen, 1997]. The effect of repeated variables is that, in some cases, the bounds on the evaluated expression will be conservatively suboptimal or too wide [Kreinovich et al., 2002]. In the continuous and discrete models of exponential growth, Equations 1 & 5, each variable appears only once, therefore interval arithmetic can be naively applied.

A statistical distribution can have uncertain moments, for example, bounds on the mean or standard deviation (c.f. [Ferson, 2002]). To propagate epistemic uncertainty through a Monte Carlo simulation interval analysis is used.

Equation 7 can be written in two *intervalized* forms

$$h(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) = \mathbf{N}_0 \exp\left(\sum_{i=1}^{T} g_i(\bar{\mathbf{r}}, \sigma_{\mathbf{r}})\right) = \mathbf{N}_T \tag{8}$$

$$j(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) = \mathbf{N}_0 \exp\left(T\bar{\mathbf{r}} + \sigma_{\mathbf{r}} \sum_{i=1}^{T} g_i(0, 1)\right) = \mathbf{N}_T \tag{9}$$

If all the parameters for Equations 8 & 9 are degenerate intervals then the two functions are equivalent given the same set of random deviates. A degenerate interval is defined as $\mathbf{X} = [\underline{x}, \overline{x}]$, where $\underline{x} = \overline{x}$. If $\mathbf{N}_0 \in \mathbb{IR}$ is the only non-degenerate parameter the expressions are still equivalent because $\mathbf{N}_0$ appears only once in each of the expressions. When $\bar{\mathbf{r}} \in \mathbb{IR}$ or $\sigma_{\mathbf{r}} \in \mathbb{IR}$ then the expressions do not give equivalent results, and it follows from subdistibutivity of interval arithmetic $j(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) \subseteq h(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T)$.

In Equation 8 the dependency between the statistical moments for the individual variates in the sum $g_1(\bar{r}, \sigma_r) + g_2(\bar{r}, \sigma_r) + \ldots + g_T(\bar{r}, \sigma_r)$ is not accounted for. The dependency occurs in that the $\bar{\mathbf{r}}$ and $\sigma_{\mathbf{r}}$ occur repeatedly in the expression as statistical moments for g. Due to the ability to factor out the mean and variance from a normal variate the sum of variates can be algebraically rearranged to take into account that $\bar{r} \in \bar{\mathbf{r}}$ and $\sigma_r \in \sigma_{\mathbf{r}}$ are fixed values:

$$\bar{r} + \sigma_r g_1(0,1) + \bar{r} + \sigma_r g_2(0,1) + \ldots + \bar{r} + \sigma_r g_T(0,1) = T\bar{r} + \sigma_r \sum_{i=1}^{T} g_i(0,1).$$

The question then becomes which of the formulations, Equations 8 or 9, is correct. The answer to this question depends on one's philosophical view of what an interval is. If the belief is that there exists a single fixed value bounded by an infinimum and supremum which bounds the uncertainty about ones estimate of the fixed value, then Equation 9 gives the optimal answer. However, if one thinks of an interval as representing a closed bounded set then there is no reason to believe that the $\bar{r}$ is fixed at each point in time. Allowing $\bar{r}$ or σ_r not to be fixed leads to widening bounds on N_T.

5 Quasi-extinction risk

The study of population viability is focused on quantifying the risk of a population falling below a critical period over a fixed time period. Rather then focusing entirely on total extinction, $N = 0$, the concept of quasi-extinction risk has been developed [Ginzburg et al., 1982]. Quasi-extinction risk is the probability that a population will fall below a given threshold during the simulation. Because intervals were used to propagate uncertainty through the simulation upper and lower bounds on the quasi-extinction risk curve must also be generated.

For Monte Carlo simulations of population dynamics the quasi-extinction decline curve is generated from the minimum of each k series of abundance.

$$N_{\min_k} = \min\left(N_{1,k}, N_{2,k} \ldots, N_{T,k}\right). \tag{10}$$

Note that the initial abundance $N_{0,k}$ is not included in the calculation of the minimum [Akçakaya et al., 1999]. For a sorted list of abundances $N_{\min_1} \leq N_{\min_2} \leq \ldots \leq N_{\min_K}$, where K is the total number of simulations, a cumulative probability $p_k = k\frac{1}{K}$ is associated with each $N_{\min_k}$.

For interval data the minimum is defined as

$$\min\left(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}\right) = \left[\min(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n), \min(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)\right] \tag{11}$$

$$\mathbf{N}_{\min_k} = \min\left(\mathbf{N}_{1,k}, \mathbf{N}_{2,k} \ldots, \mathbf{N}_{T,k}\right). \tag{12}$$

To generate the quasi-extinction decline risk curve — the cumulative distribution function of minimum abundances — for interval data the infinimum and supremum are sorted separately

$$\underline{N}_{\min_1} \leq \underline{N}_{\min_2} \leq \cdots \leq \underline{N}_{\min_K}$$
$$\overline{N}_{\min_1} \leq \overline{N}_{\min_2} \leq \ldots \leq \overline{N}_{\min_K}.$$

A probability mass $p_k = k\frac{1}{K}$ is associated with each sorted $\underline{N}_{\min_k}$ and $\overline{N}_{\min_k}$. To conservatively bound the quasi-extinction decline curve a step function is used. The bounds on the infinimum of the CDF are

$$\underline{\mathrm{CDF}}(x) = \begin{cases} \text{if } \underline{N}_{\min_1} \leq x < \underline{N}_{\min_2} \text{ then } 1/K \\ \text{if } \underline{N}_{\min_2} \leq x < \underline{N}_{\min_3} \text{ then } 2/K \\ \vdots \\ \text{if } \underline{N}_{\min_{K-1}} \leq x < \underline{N}_{\min_K} \text{ then } 1 \end{cases} \tag{13}$$

and the bounds on the supremum are

$$\overline{\mathrm{CDF}}(x) = \begin{cases} \text{if } \overline{N}_{\min_1} < x \leq \overline{N}_{\min_2} \text{ then } 1/K \\ \text{if } \overline{N}_{\min_2} < x \leq \overline{N}_{\min_3} \text{ then } 2/K \\ \vdots \\ \text{if } \overline{N}_{\min_{K-1}} < x \leq \overline{N}_{\min_K} \text{ then } 1 \end{cases}. \tag{14}$$

6 Acknowledgments

I would like to thank Dr. Scott Ferson of Applied Biomathematics for his comments on this paper, and Dr. Lev Ginzburg of Stony Brook University for his support.

References

[Akçakaya et al., 1999] Akçakaya, H. R., Burgman, M. A., and Ginzburg, L. R. (1999). *Applied Population Ecology: Principles and Computer Excercises using RAMAS Ecolab.* Sinauer Associates, Inc., Sunderland, Massachusetts.

[Ferson, 2002] Ferson, S. (2002). *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers.* Lewis Publishers, Boca Raton, Florida.

[Ginzburg et al., 1982] Ginzburg, L. R., Slobodkin, L. B., Johnson, K., and Bindman, A. G. (1982). Quasiextinction probabilities as a measure of impact on population growth. *Risk Analysis*, 21:171–181.

[Hansen, 1997] Hansen, E. (1997). Sharpness in interval computations. *Reliable Computing*, 3:7–29.

[Kreinovich et al., 2002] Kreinovich, V., Longpre, L., and Buckley, J. J. (2002). Are there efficient necessary and sufficient conditions for straightforward interval computations to be exact? In *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, pages 94–96, Toronto, Canada.

[Lewontin and Cohen, 1969] Lewontin, R. C. and Cohen, D. (1969). On population growth in a randomly varying environment. *Proceedings of the National Academy of Sciences*, 62:1056–1060.

[Moore, 1966] Moore, R. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.

[Moore, 1979] Moore, R. E. (1979). *Methods and Applications of Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.

[Rump, 1999a] Rump, S. M. (1999a). Fast and parallel interval arithmetic. *BIT*, 39(3):534–554.

[Rump, 1999b] Rump, S. M. (1999b). Intlab – interval laboratory. In Csendes, T., editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers.