Oecologia (Berl.) 32, 79–84 (1978)

*Oecologia*
© by Springer-Verlag 1978

# Choosing Category Size in a Stage Projection Matrix

John Vandermeer

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

**Summary.** A basic problem associated with choosing category size in a stage projection matrix is described. "Errors of estimation" are large if the category size is chosen too small and "errors of distribution" are large if the category size is chosen too large. An approximate technique of balancing these two error types is suggested as a solution to this dilemma.

## Introduction

The development of demographic techniques has stemmed largely from the study of human populations. As a result, almost all of demographic analysis assumes, at the start, that an overwhelming fraction of the individual variability in survivorship and fecundity is accounted for by age. Thus, individuals are partitioned into age categories and age specific survivorship and fecundity values are estimated and plugged into well known formulas (Lotka, 1926). In particular, the population projection matrix of Leslie (1945, 1948) and Lewis (1942) has become a standard tool (Goodman, 1968; Keyfitz, 1968; Keyfitz and Flieger, 1971).

In a variety of situations this approach is difficult. For example, in many biological populations the individuals cannot be aged. The standard approach is useless in such a situation. In other cases, biological constraints inhibit using the standard metods. The most obvious examples are in various plant species where some variable other than age accounts for most of the variability in survivorship and reproduction (Harper and White, 1974; Werner, 1975).

In response to problems of this sort, Lefkovitch (1965) proposed an alternative model, analytically similar to the Lewis-Leslie model (Vandermeer, 1975), but applicable to a broader spectrum of biological situations. Rather than categorizing individuals according to their age, this approach categorizes individuals according to some arbitrary morphological or physiological feature. For example, instead of defining stage one as composed of all individuals between the ages of 0–5, we define stage one as all first instar larvae, or all seedlings between 0 and 20 cm in height. The Lefkovitch approach, hereafter referred to as the stage projection matrix, is

0029-8549/78/0032/0079/$01.20

particularly useful for populations of long-lived individuals such as most perennial plant species.

The stage projection matrix has not yet been used widely. Lefkovitch (1965a) used the approach in studying the cigarette beetle, *Lasioderma serricorne*, Hartshorn (1974) applied the method to the dominant rainforest tree *Pentaclethra macroloba*, and Bosch (1971) made an abortive attempt at modeling the California redwood population with this model. The model will undoubtedly receive much more attention in future years, especially as interest builds in the population dynamics of long-lived perennials.

## The Problem

In developing a stage projection matrix, we must (1) establish stage categories and (2) estimate probabilities of transition from one stage to another over an arbitrarily selected time interval. Once we have grouped individuals into stage categories (e.g. seedlings between 0 and 15 cm) it is a simple matter to estimate transition probabilities. This is done by measuring each individual with respect to the categorization variable (height of seedling, length of body, etc.) at two points in time. Some of the individuals will die during this time interval, some will advance or regress to another stage, and some will stay in the same stage (retain the same value for the categorization variable). Estimates of transition probabilities follow directly and elementarily from such data.

But the original establishment of stage categories is not so elementary. The way in which the categories are chosen may have a significant effect on the probability estimates. In particular, the probability estimates are subject to two types of 'error'; first, a 'sample error' increasing larger as the category size decreases, and second, an 'error of distribution' increasing as the category size increases.

The first error type is of the usual sort. Given a finite number of individuals, as the category size decreases, probability estimates will be made with a smaller and smaller number of individuals. Suppose, for example, that we have 25 seedlings, 14 of which are from 1 to 5 cm tall and 11 of which are from 6 to 10 cm tall, at the begining of the time interval. If we choose the first stage category as 1 to 5 cm, the transition probabilities will be estimated with 14 individual plants. But if we choose the first stage category as 1 to 10 cm, the transition probabilities will be estimated with 25 individuals. From this point of view, obviously the larger category gives the better estimate.

The second error type is somewhat less obvious. The estimates of transition probabilities will depend in part on the distribution of individuals within the stage category. For example if the stage is defined as all seeedlings between 0 and 10 cm, and the sample includes only individuals 1 cm tall, the estimated transition probability likely will be much different from a probability calculated from a sample composed of only individuals 10 cm tall. The way individuals are distributed within the stage category will influence the probability estimate. Theoretically, the individuals within the stage category should be distributed according to the stable stage distribution (Vandermeer, 1975). The finer the category, the less likely that a skewed within-category distribution will influence

the probability estimate (theoretically, an infinitely small category size does not generate a distribution error).

Thus, if the category size is too small, the error due to sampling is potentially great; and if the category is too large, the error due to distribution is also potentially great. How then does one choose the size of categories so as to minimize, in some sense, these two errors?

## A Proposed Solution

In general, we wish to estimate the potential error incurred through sampling and through distribution. The error due to sampling should be a decreasing function of category size, while the error due to distribution should be an increasing function of category size. The point at which the two functions intersect will be called the 'optimal' category size – clearly an arbitrary optimization criterion.

Though there is no absolutely correct way of computing the two errors, because they are of such a different nature, the following is presented as a reasonable procedure for minimizing a particular manifestation of each error type.

Let $X_i(t)$ be the value of the categorization variable for the ith individual at time $t$. Define the category size as $X^*$ to $X^{**}$. Let $N_1$ be the number of individuals for which $X^* \leqq X_i(1) \leqq X^{**}$ at the beginning of the time interval, and let $M_1$ be the number of those individuals for which $X^* \leqq X_i(2) \leqq X^{**}$ at the end of the time interval. Define $\hat{P} = M_1/N_1$.

Compute the 'growth increment' for each individual as $I_i = X_i(2) - X_i(1)$. Consider only those $N_1/2$ individuals with the smallest $I_i$ (or $(N_1 - 1)/2$ if $N_1$ is odd). Let $M_2$ be the number of individuals for which $X^* \leqq X_i(2) \leqq X^{**}$, considering only those individuals with the smallest $I_i$. Define $P_s(1) = 2M_2/N_1$.

Similarly, consider only those $N_1/2$ individuals with the largest $I_i$. Let $M_3$ be the number of individuals for which $X^* \leqq X_i(2) \leqq X^{**}$, considering only those individuals with the largest $I_i$. Define $P_s(2) = 2M_3/N_1$. Finally, let the error due to sampling be,

$$\psi = (P_s(1) - \hat{P})^2 + (P_s(2) - \hat{P})^2.$$

This term is similar to a variance one would get if two samples had been tken, the first time sampling only the very slowest growing organisms and the second time only the fastest growing organisms.

Next we compute the error due to distribution. Let $M_4$ be the number of individuals for which $X^* \leqq X^* + I_i \leqq X^{**}$ and define $P_d(1) = M_4/N_1$. Similarly let $M_5$ be the number of individuals for which $X^* \leqq X^{**} + I_i \leqq X^{**}$ (note that $I_i$ will not necessarily be positive since individuals may decrease in the value of the categorization variable), and define $P_d(2) = M_5/N_1$. Let the error due to distribution be,

$$\phi = (P_d(1) - \hat{P})^2 + (P_d(2) - \hat{P})^2.$$

This term is again similar to a variance. This time, however, all individuals were changing by exactly the same amount as those individuals actually sampled, but

one sample had all individuals concentrated at the beginning of the stage category, while the other sample had them all concentrated at the end of the category.

Begin with the largest possible category size and compute $\psi$ and $\phi$. Shorten the category size and recompute $\psi$ and $\phi$. Continue shortening the category size until the number of individuals within the category is smaller than some predetermined level and plot $\psi$ and $\phi$ against category size. Draw smooth curves through the points and determine where the curves intersect. The intersection point gives the optimal category size for the first stage category. Repeat the process for the second, third, etc.... categories.

## An Example

The following example is presented to illustrate the computations involved. It is obviously an unrealistic example, the numbers being chosen for heuristic simplicity.

The first three columns of Table 1 give the raw data ($i$, $X_i(1)$, and $X_i(2)$) for 10 individuals. The $X$'s can represent any categorization variable (e.g., height of plant in cm). The fourth column gives the growth increment, $I_i$. Columns 5 through 12 give appropriate calculations for each category size (for convenience, only category sizes that include an even number of individuals were used).

For category 1–10, all but individuals 9 and 10 remained in the category, giving $\hat{p}=8/10=0.80$. Setting all individuals equal to 1 and adding the increments gives 4, 0, 4, 1, 4, 2, 3, 0, 4, 2 for individuals 1–10 respectively. All but two of these (the two zeros) are within the appropriate stage category, making $P_d(1)=8/10=0.80$. Setting all individuals equal to 10 and adding the increments gives 13, 9, 12, 10, 13, 11, 12, 9, 13, 11 for individuals 1–10 respectively. Only three (two 9s and a 10) are within the appropriate stage category, making $P_d(2)=3/10=0.30$. Thus, $\phi=(0.80-0.80)^2+(0.80-0.30)^2=0.25$. Considering the $N_1/2$ individuals with the smallest increments, we have individuals 2, 8, 4, 6, and 10, the first four of which stayed in the same stage categorization, making $P_s(1)=2M_1/N_1=2(4)/10=0.80$. Those $N_1/2$ individuals with the largest increments are individuals 3, 7, 1, 5, and 9, four of which

**Table 1.** Exemplary data for finding optimal category size (see text for explanation)

| i | $X_i(1)$ | $X_i(2)$ | $I_i$ | Cat. size | $P$ | $P_s(1)$ | $P_s(2)$ | $P_d(1)$ | $P_d(2)$ | 4 | ∅ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 3 | 1–10 | 0.80 | 0.80 | 0.80 | 0.80 | 0.30 | 0.00 | 0.25 |
| 2 | 2 | 1 | −1 | 1–7 | 0.88 | 1.00 | 0.75 | 0.75 | 0.38 | 0.03 | 0.16 |
| 3 | 3 | 6 | 3 | 1–5 | 0.50 | 0.67 | 0.33 | 0.83 | 0.33 | 0.06 | 0.14 |
| 4 | 3 | 3 | 0 | 1–3 | 0.50 | 1.00 | 0.00 | 0.75 | 0.50 | 0.50 | 0.06 |
| 5 | 4 | 7 | 3 | | | | | | | | |
| 6 | 5 | 6 | 1 | | | | | | | | |
| 7 | 7 | 9 | 2 | | | | | | | | |
| 8 | 7 | 6 | −1 | | | | | | | | |
| 9 | 8 | 11 | 3 | | | | | | | | |
| 10 | 10 | 11 | 1 | | | | | | | | |

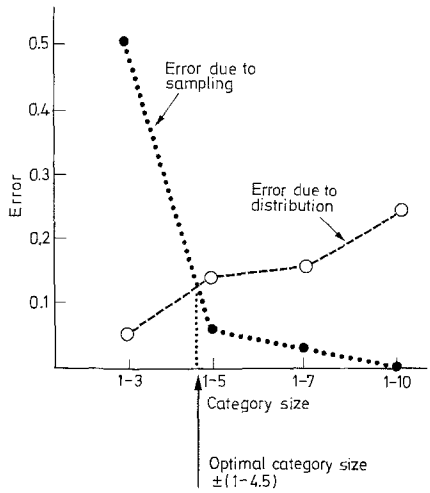


**Fig. 1.** Graph of "errors" against category sizes. Dotted line is sampling error and solid line is distribution error

stayed in the same stage, making $P_s(2)=2(4)/10=0.80$. Thus, $\psi=(0.80-0.80)^2+(0.80-0.80)^2=0$.

Consider the next smaller stage category by removing the last two individuals from the computation (individuals 9 and 10) and repeat the above.

Repeat the procedure until no fewer than three individuals are in the category (three was chosen arbitrarily – in practice the number would surely be larger), and plot $\psi$ and $\phi$ against size as has been done in Figure 1.

Once the first (smallest) category size has been chosen, the entire process is repeated for the next largest category.

## Discussion

The method proposed in this paper is based on a minimization of two error types so as to choose category sizes in which the effects of sampling and within-category skeweness will be minimal. While certainly not the only procedure which could be used, it seems reasonable and is certainly easy to apply. The method unfortunately is limited to the a posteriori analysis of data. It provides no basis for the a priori selection of category sizes. Similarly the method is not meant to apply to those situations in which stages are defined biologicaly, such as insect instars.

It is important to realize that there is no guarentee that the relationship of category size to either of the two error types will be monotonic. It is only held as a reasonablle postulate that the sampling error should decrease with category size and the distribution error should increase with category size. The example in this paper was constructed to behave well for heuristic pruposes. Especially with small samples, the behavior of the two error terms could be quite erratic, unlike this contrived example.

Undoubtedly, in any real situation a relatively large amount of data will have to be dealt with, leading to rather lengthy and tiresome calculations. A FORTRAN

program with a completely worked out real example is available from the author on request.

*Acknowledgements.* I wish to thank M. Palmer, J. Stout, and R. Westman for helpful comments on the manuscript. This research was supported by NSF grant number BMS75-01158.

## References

Bosch, C., Redwoods, A.: A population model. Science **172,** 345–349 (1971)
Goodman, L.A.: An elementary approach to the population projection matrix and to the mathematical theory of population growth. Demography **V**, 382–409 (1968)
Harper, J.L., White, J.: The demography of plants. Ann. Rev. Ecol. Syst. **5**, 419–463 (1972)
Hartshorn, G.S.: The ecological life history and population dynamics of *Pentaclethra macroloba*, a tropical wet forest dominant and *Stryphnodendron excelsum*, an occasional associate. Ph. D. Thesis, Univ. of Wash., 118pp. (1972)
Keyfitz, N.: Introduction to the mathematics of population. Reading, Mass: Addison-Wesley 1968
Keyfitz, N., Flieger, W.: Population. Facts and methods of demography, 613 pp. San Franscisco: Freeman 1971
Lefkovitch, L.P.: The study of population growth in organisms grouped by stages, Biometrics **21**, 1–18 (1965)
Lefkovitch, L.P.: The effects of adult emigration on populations of *Lasioderma serricorne*. Oikos **15**, 200–210 (1965a)
Leslie, P.H.: The use of matrices in certain population mathematics. Biometrika **33**, 183–212 (1945)
Leslie, P.H.: Some further notes on the use of matrices in population mathematics. Biometrika **35,** 213–245 (1948)
Lewis, E.G.: On the generation and growth of a population. Sankhya **6,** 93–96 (1942)
Lotka, A.J.: Elements of mathematical biology. Baltimore: Williams and Willkins, 1925 (Dover reprint, 1956)
Vandermeer, J.: On the construction of the population projection matrix for a population grouped in unequal stages. Biometrics **31,** 239–242 (1975)
Werner, P.A.: Predictions of fate from rosette size in teasel (*Dipsacus fullonum* L.). Oecologia (Berl.) **20**, 197–201 (1975)

*Received July 14, 1977*