A generalized algorithm for determining category size

Kirk A. Moloney
Department of Botany, Duke University, Durham, NC 27706, USA

Summary. A revision and extension of Vandermeer's algorithm for choosing size categories from demographic data is presented along with an example of its use. The extension permits an exact consideration of sample populations which may have different underlying transition probabilities at different census periods and/or within different subpopulations. Once categories are chosen transition matrices may be computed to investigate the dynamics of the population.

Transition matrix models represent a powerful technique for obtaining information concerning the dynamic structure of a population and for predicting future population states (see Charlesworth 1980 and Keyfitz 1968 for thorough reviews). Traditionally, individuals are grouped according to age classes. Age-specific fecundity and mortality rates are estimated by sampling the population over a set time interval and a matrix of transition probabilities (the Leslie matrix) describing the life history of the organism is obtained. Only two types of transition are allowed between age classes with the Leslie model: i) production of offspring by mature individuals, and ii) transition into the next oldest age class by all individuals surviving from time $t$ to time $t+1$. A large body of theory has developed for analyzing the Leslie transition matrix model, including extensions into the theory of population genetics (Charlesworth 1980); however, the model is inadequate for many situations encountered by researchers interested in the demographic structure of populations.

The age structure of a population can be hard to determine or can often be a poor predictor of future population states, e.g. in organisms with complex life histories or in organisms of indeterminate growth (Lefkovitch 1965, Vandermeer 1978, Kirkpatrick 1984). In these cases, between-developmental-stage or between-size-class transition probabilities of individuals may be better predictors of future population states than are between-age-class transition probabilities. Lefkovitch (1965) proposed a demographic model based on the developmental stages of insects that was more general than the Leslie model, but similar analytically (Vandermeer 1975). Other researchers have used a variation of the Lefkovitch model, grouping populations according to size classes (see Werner and Caswell 1977, Hartshorn 1972, Bierzychudeck 1982, and Meagher 1982 for examples). The Lefkovitch model permits a greater range of transition classes, including changes from larger size classes to smaller ones. For some organisms, such as non-woody perennial plants, this is essential for the development of a realistic model.

It is not necessarily a simple matter to develop a classification scheme for assigning individuals to size or stage categories, especially when there is continuous variation for the classification character within a population. Vandermeer (1978) proposed an algorithm for choosing appropriate size or stage categories from demographic data. His strategy was to maximize the information content of the Lefkovitch matrix by minimizing two types of negatively correlated error. The first type of error occurs if classification categories are made too large. In this case an underlying assumption of the transition matrix model, that individuals falling within the same category have the same transition probabilities, is violated. Larger individuals will most likely have transition probabilities that differ from those of smaller individuals, both as a function of initial size and of growth rates. The estimates of transition probabilities, if a category is too large, will thus be inaccurate due to what Vandermeer has termed "distribution error". In contrast, classification categories which are made too small will contain few individuals. The resulting estimates of transition probabilities will be inaccurate due to sample error. Vandermeer proposed that by choosing classification categories that minimize the effects of these two sources of error the best estimates of transition probabilities can be obtained.

A quantitative method for estimating the sample and distribution error associated with any arbitrarily chosen classification category was first presented by Vandermeer in 1978. By systematically determining these two types of error for a range of category sizes, categories for a transition matrix analysis can be determined. However, as acknowledged by Vandermeer in his paper, the definitions he proposed for both distribution and sample error were somewhat arbitrary. Also, Vandermeer's algorithm is only appropriate for choosing category sizes from a demographically homogeneous population sampled over a single time interval. No allowances are made for differences among subpopulations or between censuses. If it is of interest to compare the demographic processes of populations affected by different environmental or genetic backgrounds or to compare these processes at different stages in the establishment of a population, the classification algorithm must be extended to accommodate differences in transition probabilities associated with subsets of the sample population.

This paper redefines sample and distribution error and extends Vandermeer's classification algorithm to account...
for differences in transition probabilities among subpopulations and census periods. Although the general development of the algorithm presented closely parallels the development of the original by Vandermeer, the notation has been revised to allow the redefinition of both types of error and to allow the extension to more general cases. An example of the algorithm's use is also given.

The general algorithm

Preliminary development. The first step in determining appropriate classification categories is to obtain an estimate \( \hat{P} \) of the true probability \( P \) that an individual starting in a category at census \( j \) will be in the same category at census \( j + 1 \). The maximum size allowed within a category for each calculation will be represented as \( M_{\text{max}} \) and the minimum as \( M_{\text{min}} \). Let \( m_k(j) \) be the size of individual \( X_k \) at census \( j \) in subpopulation \( k \), with \( n_j \) and \( n_k \) the number of censuses and subpopulations, respectively. Change in the size of \( X_k \) from census \( j \) to census \( j + 1 \) will be denoted as \( d_k(j) \).

\[
d_k(j) = m_k(j+1) - m_k(j)
\] (1)

Summation over the range of either \( i \) or \( k \) will be represented by a dot (.) replacing the subscript. All summations over \( j \) will be from \( 1 \) to \( n_j - 1 \).

Two dummy variables, \( t \) and \( r \), are used in calculating \( \hat{P} \).

If \( M_{\text{min}} \leq m_k(j) \leq M_{\text{max}} \)

\[ t_k(j) = 1, \quad \text{else } t_k(j) = 0 \] (2)

and,

if \( M_{\text{min}} \leq m_k(j) \) and \( m_k(j+1) \leq M_{\text{max}} \)

\[ r_k(j) = 1, \quad \text{else } r_k(j) = 0 \] (3)

Summation of \( t \) over \( i \) and \( k \) provides a total of the number of individuals that start within the classification category delimited by \( M_{\text{min}} \) and \( M_{\text{max}} \) at census \( j \), whereas summation of \( r \) provides a total of the number of individuals that are within the category at both censuses \( j \) and \( j + 1 \). \( \hat{P} \) can be calculated using the dummy variables just defined.

\[
\hat{P} = \sum \frac{t_j}{n_j} \sum \frac{r_j}{n_j}.
\] (4)

The estimated number of individuals leaving the interval \( M_{\text{min}} \) to \( M_{\text{max}} \), \( \hat{Q} \), is equal to \( 1 - \hat{P} \).

Distribution and sample error. The demographic behavior of an individual in a size class model, ignoring reproduction, can be characterized by two parameters: i) size at the beginning of a census period, \( m_k(j) \); and ii) change in size during a census period, \( d_k(j) \). This can be represented as a vector quantity in one dimension (Fig. 1a). In the special case of death by an individual we will assume \( m_k(j+1) = 0 \). The transition probabilities assigned to an individual occurring in a particular size category are the same as those of any other individual falling within the same category at the beginning of the census period. It is equivalent to setting the initial size of all individuals \( X_k \), with \( t_k(j) = 1 \) for the interval \( M_{\text{min}} \) to \( M_{\text{max}} \) at census \( j \), equal to \( m_k(j) = (M_{\text{min}} + M_{\text{max}}) / 2 \) and further assuming that calculations of \( \hat{Q} \) and \( \hat{P} \) are unaffected by substituting \( m_k(j+1) = m_k(j) + d_k(j) \) for \( m_k(j+1) = 0 \). Equation 3 (Fig. 1b).

The degree to which this assumption is violated will be defined as Distribution Error (DE).

Fig. 1A, B. Distribution of vectors representing a hypothetical population sample with values of \( m_k(j) \) falling between boundaries of \( M_{\text{min}} \) and \( M_{\text{max}} \). A Vectors with tails at \( m_k(j) \) and heads at \( m_k(j) + d_k(j) \). B Vectors with tails at \( m_{\text{max}}(j) \) and heads at \( m_{\text{max}}(j) + d_k(j) \) for the same sample population as in A

A new dummy variable, \( r_k^*(j) \), is used in calculating the DE for any arbitrary category size \( M_{\text{min}} \) to \( M_{\text{max}} \).

If \( M_{\text{min}} \leq m_k(j) \) and \( m_k(j+1) \leq M_{\text{max}} \)

\[ r_k^*(j) = 1, \quad \text{else } r_k^*(j) = 0 \] (5)

An estimate of \( P \) accounting for DE can be calculated for each of the temporal and spatial subdivisions of the population as \( P_{\text{st}}(j) \).

\[
P_{\text{st}}(j) = r_k^*(j) / t_k(j).
\] (6)

DE can then be estimated as a variance using 6 as a measure of the deviation of the sample population from the expected value \( \hat{P} \).

Several secondary considerations must be included in any definition of DE. First, since transition elements contain information associated with both remaining in size categories and with leaving them over a census period, deviations of both \( P^* \) and \( Q^* \) from the expected values, \( \hat{P} \) and \( \hat{Q} \), must be considered. Also, since it is as important to accurately predict small values of \( P \) and \( Q \) as it is to predict intermediate and large values, some standardized method for estimating variances over the interval (0,1) must be incorporated into any reasonable definition of DE. The squared ratio of the deviation \( (P^* - \hat{P}) \) to the expected value \( \hat{P} \) provides an appropriate measure (a proportional deviation from the expected value) as errors in specifying transition elements of a matrix propagate geometrically through the model with repeated multiplication.

Given the above considerations, DE can be defined as:

\[
\text{DE} = \left( (n_j - 1) n_k \right)^{-1} \sum \left[ (1/2)(P_{\text{st}}(j) - \hat{P})/\hat{P} \right]^2 + 1/2 \left( Q_{\text{st}}(j) - \hat{Q} \right)/\hat{Q} \right]^2)
\] (7)

which with

\[
C = \frac{(Q^2 + \hat{P}^2)}{2(n_j - 1) n_k} \{P^* \hat{Q} \}^2
\] (8)

can be simplified to

\[
\text{DE} = C \sum \{P_{\text{st}}(j) - \hat{P} \}^2.
\]
Ideally, category sizes should be chosen so as to minimize DE. Ignoring the effect on the calculation of subdividing the population, DE will approach a value of zero as \( M_{\text{max}} - M_{\text{min}} \) approaches zero since there will be little or no difference between \( m^* \) and \( m \). If DE were the only criterion used in choosing category sizes we would always choose the smallest category size possible and have no need of a mathematical technique. However, this is inappropriate as the sample sizes used in estimating transition probabilities decline as category sizes are made smaller. The estimates of \( P \) and \( Q \) will thus become increasingly inaccurate. The error introduced by this inaccuracy will be referred to as SE.

An estimate of sample error can be obtained using a methodology similar to the one devised for DE. For any subdivision \( k \) of the population at census \( j \) consider only the subset of individuals for which \( t_{ik}(j) = 1 \). An estimate can be made of SE by resampling from this subset with replacement until obtaining a new subset, \( X_{ik}^*(j) \), which contains the same number of individuals. The subscript \( s \) will refer to the individual resampling events within each subset. Individuals \( X_{ik}^*(j) \) are of size \( m_{ik}^*(j) \) at census \( j \) with \( a_{ik}^*(j) \) defined as in equation 1. Each subset is resampled \( n_i \) times to increase the accuracy of the estimate of SE.

A new dummy variable \( r_{ik}^*(j) \) is defined as follows:

\[
\text{If } M_{\text{min}} \leq m_{ik}^*(j) \text{ and } n_{ik}^*(j+1) \leq M_{\text{max}} \text{ then } r_{ik}^*(j) = 1,
\]

\[
\text{else } r_{ik}^*(j) = 0.
\]

An estimate of \( P \) accounting for SE, \( P_{ik}^*(j) \), can be calculated for each resampling as

\[
P_{ik}^*(j) = \frac{r_{ik}^*(j)/t_{ik}(j)}{P}
\]

The definition of SE follows directly, using the same arguments as in the derivation of equation 9.

\[
\text{SE} = C^{**} \sum_j \sum_k \{ P_{ik}^*(j) - \bar{P} \}^2.
\]

In this case

\[
C^{**} = \frac{\sigma_p^2 + \bar{P}^2}{2(n_i - 1)n_i \bar{P}^2}
\]

is the correct multiplier as it adjusts for the resampling events conducted in each subset.

**Determination of category size.** Category sizes are determined through a systematic analysis of SE and DE over a range of boundary values. \( M_{\text{min}} \) and \( M_{\text{max}} \). \( M_{\text{min}} \) is initially set to correspond to the smallest size allowed in the first category (defined as the category containing the smallest individuals in the population) and \( M_{\text{max}} \) to a value much greater than the expected maximum size of the category. DE and SE are then computed, \( M_{\text{max}} \) in incremented downward in steps with the two types of error being determined at each step. The upper bound to the category is set equal to the value of \( M_{\text{max}} \) at the point where the sum of DE and SE is at a minimum. (This can be determined graphically as in Fig. 2.) After the upper and lower bounds to the first category are set, \( M_{\text{min}} \) is reset to be just greater than the upper bound previously determined. The procedure is repeated until all stage categories are defined. The last category is set by either requiring a minimum number

of observations within each category or by running the algorithm though the entire data set.

**An example**

I have determined category sizes for a population of *Dantonia sericea* Nutt., a tussock forming grass species, subdivided at several levels. Five 0.25 m² quadrats were located at random at each of four positions along an elevational gradient in a mown field in Durham, NC, USA (see Fowler and Antonovics 1981b for a detailed description of the field). All individuals of *Dantonia* located within each of the quadrats were censused for number of leaves per individual (size) on a bimonthly basis between June 1983 and June 1984 (except for 12/83). The population structure of *Dantonia* differed among the four locations in the field and also to a lesser degree among the five quadrats at each location (unpublished data). Depending upon the nature of the investigation, a demographic analysis might involve
comparisons among several possible subdivisions within the sample population. The study population was considered as either one continuous population, 4 subpopulations, or 20 subpopulations at each of 3 levels of sampling intensity. The sampling levels included censuses from the months of i) 6/83 and 6/84; ii) 6/83, 10/83 and 6/84; or iii) 6/83, 8/83, 10/83, 2/84, 4/84 and 6/84.

The magnitude of sample and distribution error for any particular census schedule increases with division of the population into a greater number of subpopulations (Fig. 2). This is due to a decrease in the sample size per subdivision. There is no systematic difference in the number of categories chosen by the algorithm for 1, 4 or 20 subpopulations among the three census schedules (Table 1). However, there are differences in the number of leaves at which breaks occur, with an indication that categories at smaller leaf sizes are chosen by the algorithm to be somewhat larger with the inclusion of a greater number of subpopulations.

The introduction of intermediate census dates over one year (June to June) has some effect on the choice of size categories. The second, third and fourth categories specified by the algorithm begin at smaller leaf numbers when either one or four intermediate census dates are included in the analysis. Increasing the number of June censuses would probably not have the same effect. Part of the reason for this can be attributed to a sinusoidal change in plant size during the year. *Danthonia*, although evergreen, grows to larger size during fall and spring and decreases in size during the summer after flowering in June (unpublished data). In determining category size using data from the 6/83 and 6/84 censuses, the classification algorithm must only accommodate an annual change in size at a time when the plants are largest. With inclusion of intermediate census dates, category sizes must be assigned to accommodate more radical changes (both upward and downward) in size and substantial differences in initial size distributions at the beginning of census periods. This effect is largest in the smaller size categories which are much more susceptible to seasonal changes in the environment (unpublished data).

**Table 1.** Estimates of the number of stage categories and their base values (i.e., fewest number of leaves per plant allowed within a category), for one data base subdivided into 1, 4 or 20 subpopulations ($n_i$) and censused on 2, 3, or 6 dates ($n_j$) over a one year period

<table>
<thead>
<tr>
<th>$n_j$</th>
<th>$n_i$</th>
<th>$\sum_{j} n_{i,j}$</th>
<th>Number of stage categories</th>
<th>Estimated base values for categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>1; 6; 14; 23; 60; 118</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1; 6; 14; 30; 80</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1; 9; 15; 30; 72; 119</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1; 5; 8; 17; 36; 77</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1; 7; 12; 18; 40; 73</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>1</td>
<td>1; 7; 13; 23; 39; 62; 96</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1; 8; 20; 35; 91</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1; 13; 25; 40; 58</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>1</td>
<td>1; 12; 16; 33; 50; 75</td>
<td></td>
</tr>
</tbody>
</table>

*See text for the census dates and a complete explanation
*The last category at each level was constrained to contain at least 0.02 x $\sum_{j} n_{i,j}$ individuals

**Discussion**

Many studies have suggested that the demographic parameters of a population may vary considerably over a small spatial (e.g. Snaydon and Davies 1972; Keddy 1980; Fowler and Antonovics 1981a; Fowler 1984) or temporal (e.g. Sarukhan and Harper 1973; Weins 1976; Keddy 1980) scale. The revision of Vandermeer’s algorithm presented here allows the appropriate categorization of a demographically heterogeneous population into size classes by adjusting for variation in transition probabilities both among subpopulations and among census periods. As in the example of *Danthonia sericea*, the size classes specified by the algorithm can vary, depending upon the subdivisions imposed on the population. In general, the effect of dividing the population into smaller subunits without increasing sample size will be to increase sample error. This should lead to an eventual decrease in the number of categories specified. With *Danthonia* the sample sizes were relatively large and only small, non-systematic changes in the number of categories chosen by the algorithm were observed. There was, however, some increase in the range of leaf sizes encompassed by the smaller size categories with the introduction of greater subdivision. This could indicate a trend towards the selection of fewer categories. In contrast, the effect of changing the length or number of census periods or of including subpopulations with different demographics depends directly upon the natural history of the organism. With *Danthonia* a decrease in the size of classification categories was seen in response to the introduction of intermediate census dates. This can be ascribed, in part, to differences in transition probabilities at different times of the year.

It is critical to choose the appropriate classification categories in designing a demographic study based on size or stage classes, otherwise the results of any subsequent analyses might be weakened. As can be seen in the *Danthonia* example, this includes accounting for demographic variation among subpopulations and among census periods. The classification algorithm presented in this paper is designed to account for both types of variation in demographic parameters. Once suitable categories are chosen, transition and probabilities corresponding to specific subpopulations and census periods can be calculated. The LeKolvitch model (or difference equations based upon size classes) may then be used to investigate the effect on population processes of spatial or temporal differences in demographic parameters. This could take the form of comparative eigen-analysis, simulation studies or direct comparisons of transition probabilities among subpopulations or among census periods. Detailed analyses of small scale variation in demographic processes such as these will yield useful insights into the processes regulating the distribution of species.

**Acknowledgments.** I would like to thank J. Antonovics, D. DeSteven, R. Harris, L. Lyons, T. Meagher, B. Saunders, W. Schlessinger and the anonymous reviewers for critical comments leading to much improvement on an earlier draft. This research has been supported in part by a Grant-in-Aid from Sigma Xi to the author.

**References**


Keyfitz N (1968) Introduction to the Mathematics of Populations. Addison-Wesley, Reading, Mass
Vandermeer J (1978) Choosing category size in a stage projection matrix. Oecologia (Berlin) 32:79–84

Received January 22, 1985