What Can the Age Composition of the Population Tell Us about the Age Composition of Migrants?

Jani S. Little
Andrei Rogers

Abstract

Preliminary findings show that the age structure of a population can be useful in estimating the age composition of outmigrants. Demographers have always known that population pyramids reflect the fertility, mortality and migration processes of a region. This research, on the other hand, uses the age composition of a population to forecast the profile of outmigration. This work is motivated in part by changes in U.S. Census survey strategies that present new challenges for measuring migration. Initial investigation (presented in the attached paper) was based on state populations in 1995 and the schedules of age-specific outmigration between 1995 and 2000. Ongoing work extends these analyses using the population structure to categorize the shape of the outmigration profile (monotonically decreasing through the later years, peaking in the retirement years, or increasing for the oldest ages). Other extensions examine the generalizability of the state findings for smaller geographic units (MSAs and counties).

I. Introduction

The age breakdown of a migrating population is fundamental for demographers and policy analysts who want to make accurate population projections. In the U.S., techniques that estimate age distributions of migrants have received new interest since census taking strategies were revised. Traditionally, accurate age-specific migration flow estimates for counties, metropolitan areas, and states were made possible by the large sample who responded to the "Where did you live 5 years ago?" question on the decennial U.S. census long-form questionnaire. Replacement of the decadal long-form questionnaire with the American Community Survey (ACS) was motivated by the lower cost of administering the ACS and by the advantages of more accurate population counts during intercensal years. However, some attributes of ACS chart unknown territory that compel demographers to develop independent methods for measuring migration that might then be used to inform the use of the ACS for this purpose. One attribute of ACS is that it uses a smaller sample size than the decadal long-form sample. As a consequence, ACS migration data disaggregated by age may provide unstable and unreliable estimates of age-specific migration even for the larger geographic units. To overcome the problems of the smaller sample size, the ACS will rely on data that are averaged over five years. These will be inherently imprecise for annual estimates as well as age-specific estimates. Since the ACS migration question asks about residence last year, five-year averages will result in data that do not clearly delineate the calendar year or the age category.

In response to these challenges several indirect estimation approaches have been suggested. The most straightforward of these is the residual method, which estimates age-specific net migration by subtracting the observed population at each age from the projected age-specific population, based on assumptions of mortality and fertility (Pittenger, Castro). Another method infers the age structure of migrants from aggregate interregional flows by statistically imposing the age breakdown observed in a previous period (Rogers, A, Willekens, F, Raymer, J, 2003). Yet another method infers age-specific migration rates of older populations from available migration data of those under five years of age (Rogers and Anselmi, 2004);

The method proposed in this paper adds to the existing body of work and supplements our understanding of how auxiliary data can be used to fill in the gaps caused by missing or sparse migration data. Central to the method is the age profile of a population, which is often used as evidence of the historical fertility and mortality patterns that give rise to the population structure. For example, Figure 1 shows the contrast in age pyramids of populations in Mexico in 1970 and Sweden in 1974. The Mexico population pyramid suggests high rates of natural increase and mortality increasing fairly rapidly with age. Sweden demonstrates a population pyramid that is more typical of low rates of natural increase and morality rates that don't increase appreciably until after age 60. For the same time period, Figure 2 reveals the age profiles of internal migration for Mexico and Sweden, and it is clear that the two population structures give rise to very different migration schedules. Can the age pyramid of the population be linked to the age profile of migrants in a general way, one that is useful

regardless of place or time? In this paper we offer a foundation for how the population structure may influence the migration schedule, and we establish a predictive model that is the essence of the method for indirectly estimating the migration profile. We use the age-specific proportions of the total out-migration flows for the 50 states and the District of Colombia between 1995 and 2000 as the observed data, and the method is designed to predict the age composition of migrants leaving an area during a five-year period on the basis of the population age structure at the beginning of the period. We describe the method, which uses commonly available data and simple measures of population composition to predict the age composition of the out-migrating population.

It is well documented that profiles of age-specific migration have similar shape that is consistent over space and time, and that the profiles can be precisely represented by the Rogers-Castro (1981) model schedule, which is a smooth parameterized multi-exponential function. Initially the 51 observed profiles are fitted to the 7-parameter Rogers-Castro model schedule, and then regression models are constructed to predict the variation in each of the 7 parameters of the schedules using measures of the age structure of the state population as the explanatory variables.

The regression models are calibrated to the predict the seven parameters and from the predicted parameters a migration schedule is estimated for each of the states. The observed and the predicted migration model schedules for each of the states are compared and various measures of goodness of fit are evaluated. In addition, the effectiveness of the regression approach is compared to the effectiveness of the "standard" model schedule, which is simply the schedule generated from the most common values of the 7 parameters as documented in previous research. The standard schedule does not use any

information about the population composition, and can be considered the simplest predictive approach.

Of the sections that follow, section II details the construction of the observed data and the procedures used for calculating the migration proportion schedules. Section III describes the 7-parameter Rogers-Castro model migration schedule (Rogers, Castro, 1981) and its documented characteristics. The framework that specifies the relationships between model schedule parameters and measures of population age structure is developed in section IV. And in Section V, the regression prediction equations are estimated and the results are reported and evaluated. The final section presents conclusions about the method, and a discussion of the viability of the method, its shortcomings, and planned extensions.

II. The Observed Outmigration Data

The age-specific out-migration data for the states came from the Census 2000 Migration DVD provided by the US Census Bureau. It gives counts of persons who changed their state of residence between 1995 and 2000 and who lived to be counted in 2000. Based on a person's age in 2000, these counts are disaggregated into five-year age categories, beginning at age 5 and ending at age 85 or older, i.e. 5-9, 10-14, 15-19,….80-84, 85+. From these data we backcasted to get the numbers of outmigrants from each of the 50 states by age category in 1995, i.e. 0-4, 5-9, ….80+. These counts of migrants within each age category, who survived to 2000, are divided by the total number of persons who left that state after 1995 and survived to be counted in a different state in 2000. These are called the observed age-specific proportions or formally N(x to x+4)

where (x to x+4) is the age group which begins with age x, and N(x to x+4) is the proportion of the total number of outmigrants in the (x to x+4) age group in 1995.

There are 17 observed N(x to x+4)s for each outmigration schedule. The profiles of these data points are not smooth and consequently cannot be accurately represented by a model schedule. Smoothing is a necessity, and this begins by converting the N(x to x+4), where (x to x+4) represents a 5-year age interval, to five N(x)s that represent the proportion of migrants associated with single year ages. The first step was to assign the N(x to x+4) values divided by 5 to N(x+2), the proportion associated with age (x+2). For example, for the age interval 0-4, N(0 to 0+4)/5 is assigned to N(2), the proportion associated with age 2. Likewise, N(5 to 5+9)/5 is assigned to N(7). The next step was to use cublic spline interpolation to smooth between the values of N(2), N(7),….N(82) and to get estimated N(x)s for all other single age groups. (Advanced Systems and Design add-on to Excel.) After estimating the proportions for one-year age groups, the N(x) values were recalibrated so that the N(x)s for each state summed to 1. (Need to ask Lisa abou this…). The smoothing process resulted is the proportion of total outmigrants in each single year age category, N(0), N(1),….N(84).

III. The Rogers-Castro Model Migration Schedule

Migration proportion schedules universally exhibit a common shape, and decades of research have shown that these profiles can be accurately represented with a multiexponential function, called the Rogers-Castro model schedule. From previous work by Rogers, Castro and others, we have learned much about this model migration

schedule, including how to interpret the parameters and the expected values ranges for each of the parameters. Figure 3 displays a typical migration model schedule and its 7-parameters. Starting with relatively high levels during the early childhood ages, the "infant peak" is captured by the parameter a1. Then the proportions decrease monotonically to a low point around age 10. The rate of decrease is represented by the alpha1 parameter, also referred to as the pre-labor force slope. This is followed by an increasing slope (lambda1) eventually reaching a peak (usually called the labor peak) between ages 15 and 22. The peak is captured by the a2 parameter. The schedule then decreases once again to the ages of retirement at a rate represented by the alpha2 parameter, called the post labor force slope. Finally the schedule levels off around some constant level (c). Sometimes a post-labor force component appears showing a bell-shaped curve that represents the raised prevalence of migration during the retirement years, or other times there is an upward slope that increases monotonically to the last age included in the schedule. These post-labor force components require extensions to model schedule from 7 parameters to 9 or 11 parameters. Given the purposes of this paper, the additional complexity of 9 or 11 parameters was not warranted.

Castro and Rogers (1983) demonstrated that the shape of the migration profile can provide a wealth of information that can be used to describe the characteristics of the migrating population, such as, "Is it male or female dominate?", or "Is the pre labor force migration reflective of low or high family dependency patterns?" For the purposes of this paper, the smoothed observed data values for each profile are fitted to this model because it parsimoniously and accurately represents each profile and because it reduces the 85 data values in a migration profile down to the 7 parameters of the model schedule.

The fitting process was done the nonlinear regression procedure in SPSS. The average R^2 value generated from correlating the smoothed data with the data implied by the fitted model schedule is .98 across the 51 states. R^2 values ranged from 0.95 for Maine (Figure 4a) to 0.99 for Pennsylvania (Figure 4b).

The variation in the parameters can be summarized in Table1 and Figure 5.

Table 1. Summary of Parameter Variation

	alpha1	a1	alpha2	a2	mu2	lambda2	c
average	0.0169	0.0484	0.0448	0.0738	17.2394	0.2354	0.0011
standard deviation	0.0027	0.0158	0.0085	0.0137	2.6492	0.0873	0.0004
maximum	0.0219	0.0952	0.0664	0.1000	23.6171	0.4104	0.0034
minimum	0.0105	0.0220	0.0297	0.0501	15.0000	0.1000	0.0010
coefficient of variation	0.1624	0.3256	0.1899	0.1859	0.1537	0.3707	0.3459

III. Linking Population Composition with Outmigration Profiles

The foundation for linking population profiles with migration schedules is guided by the general form of the Rogers-Castro migration model schedule and the interpretations of the parameters; by the scant literature that documents how population composition can influence migration profiles; and by well reasoned arguments and relationships suggested by the data.

Predicting migration schedules presents challenges that stem from the fact that they are not captured by a single variable, but by a mathematical function, which is defined uniquely by 7 parameters. To add to the complexity, the parameters are not equally important in determining the shape of the profile, and some parameters are highly correlated with others. (See Table 2.) One parameter having a relatively large value may

necessarily restrict the range of another parameter value. This is intuitive since that the sum of all age-specific proportions of the migrating population must total 1. Therefore, a large contribution at one age will necessarily reduce the contributions of other ages.

Another guiding principal comes from the general shape of the Rogers-Castro model schedule. This gives us rough guidelines for the ages with the highest propensities for migration and the relative size of the age groups in the population will effect the representation among migrants.

To understand these dependencies, a preliminary principal components analysis was done to partition the parameters of the model schedules into components, or subsets of parameters that are intercorrelated, yet the subsets are uncorrelated with each other. 89% of the variance in the 7 parameters of the 51 schedules are represented by three principal components. These results are reported in Table 3. The first component accounts for 49% of the variance and is a composite of the child migration parameters (a1, alpha1), the labor slope (lambda1), and the age of peak migration (mu2). This component will be referred to as representing "EarlyYears Migration." The second component accounts for 21% of the variance and is comprised of two parameters, a2 (the height of the career peak) and alpha2 (the post labor slope). This component is named the "Middle Years Migration." The third component accounts for 17% of the variance and is represented by parameter c (the minimum value of the schedule) and is called "LateYears Migration" .

Table 2. Correlations of Parameters

	Alpha1	a1	alpha2	a2	mu2	lambda2	c
alpha1	1.0						
a1	.639	1.0					
alpha2	.303	.477	1.0				
a2	.080	-.104	.717	1.0			
mu2	.534	.677	.661	.334	1.0		
Lambda2	-.441	-.724	-.422	.021	-.779	1.0	
C	-.202	.103	-.015	.132	.311	-.206	1.0

Table 3. Principal Component Analysis, Rotated Component Matrix

	Component		
	1	2	3
a1	.737	.104	-.430
al1	.926	.009	-.013
lam2	-.872	-.053	-.220
a2	-.098	.969	.069
mu2	.808	.424	.252
al2	.431	.845	-.056
C	.099	.043	.952

The separation suggested by the principal components analysis is illustrated in Figure 6. And since the contribution to total variance is greatest for the first component we begin by drawing linkages between population composition and the parameters represented in the first principal component.

Early Years Migration

One principal documented in the literature is that migration is age selective. Among adults, young adults are the most mobile group in any population (Castro and Rogers, 1983), and in some special areas people show high migration propensities during the retirement years. Among children in the dependent years, infants are the most mobile, and for them migration must occur within the family unit, suggesting that families with infants are more inclined to migrate than families with older children. Variation in the infant peak (a1) parameter is likely to be explained in part by the infant dependency ratio, i.e the ratio of the number of infants (ages 0-4) divided by the population in the migrating years (15-34). This measure is similar to a fertility rate, but the total number of men and women in the ages of peak adult migration is used as the base population. Thus, we expect that for populations exhibiting high levels of natural increase, infants will be a larger component of the migration profile because family migration will be more prevalent. Castro and Rogers (1983) confirmed this with sensitivity analyses and concluded that the shape of the migration proportion schedule is highly sensitive to changes in the dependency level.

A second hypothesis is that the infant peak is influenced by the age distribution of the heads of households during the migration years. This principal was reported by Castro and Rogers (1983) when they found through sensitivity analyses that the N(x) schedule is very sensitive to changes in the age distribution of family heads. Based on Figures 7a and 7b, the height of the infant peak is likely to be positively related to the proportion of the population in the peak years of migration and in the beginning years of family formation, i.e. ages 20-24. The three states show dramatic differences in the infant peak (Figure 7a), with Utah having the highest level, then Delaware and the lowest

level is in Maine. In Figure 7b the population distributions show Utah has the highest proportion in group 20-24 (.086) followed by Delaware (.067) and Vermont (.064).

The alpha1 parameter captures the prelabor slope, and the larger the value the steeper the descent, which suggests more younger children than older children are migrating with their parents. From Figures 7a and 7b it is apparent that this slope seems to be related to the size of the youngest group (ages 0-4) as compared to older group (ages 5-9). This is called the infant-youth ratio. For example, Delaware has the steepest slope (.095) and at the same time the infant-youth ratio is 1.02.

Lambda2 is the labor slope parameter and a larger value implies a faster ascent to the labor peak. Two hypotheses came from examining Figures 8a and 8b where West Virginia has the steepest slope (.41), followed by Nebraska (.25) and Delaware (.15). First, it appears that the age of the largest cohort in the career migration years (15 to 34) affects the labor slope. The younger the largest cohort, the earlier the beginning labor migration and the steeper the slope. In contrast, if the largest cohort is older the labor migration will be more dispersed over the age groups and the slope will be flatter. West Virginia, for example, has its largest cohort at age 18 versus Nebraska and Delaware where the age of largest cohort is 34.

A final hypothesis about the early migration profile parameters has to do with mu2, the peak age of labor migration. Washington D.C. and Maine are examples of two extremes (Figures 9a and 9b). Washington D.C. has the older average age (mu2=20) and Maine has the younger average age (mu2=15). From the population distributions in Figure 9b, the most striking difference between these two population distributions is the difference between the size of the cohort in ages 15-19 and the size of the group ages 20-

24. In Washington D.C. the older group is much larger than the younger group (40,899 vs 25,915) suggesting the large older cohort will influence the migration profile by shifting the migration peak to the right. In Maine, the younger cohort is larger than the older cohort (85,515 vs 80.002) and this difference will shift the migration peak to the left.

Middle Years Migration

Based on the principal components analysis the parameter that represents the height of the career migration peak (a2) and the parameter that measures the post labor slope (alpha2) are highly correlated and are distinct from the other parameters in their contribution to the model schedule. The parameter a2 is highest when the peak is most pointed suggesting that career migration is more concentrated at a single age or a cluster of ages. This is unlikely if the population in the ages with the highest propensities for labor migration is dominated by the older cohorts. This is suggested in the Figures 10a and 10b. North Dakota shows a peaked profile and Nevada has a much flatter profile. The North Dakota population is more uniformly distributed through the years 20-24, 25-29 and 30-34 (cohort sizes are 46,113; 43,394; 46,083). In contrast, in Nevada there is a relatively large population in the later years of career migration. In fact, the population increases during these years (cohort sizes are 92,170; 112.909; 140,350 respectively), which suggests higher proportions are migrating throughout these years and a flatter labor migration pea results.

The other middle year migration parameter is alpha2. It represents the decreasing post labor slope and the larger values indicate a more rapid descent. Figure 11a demonstrates that Arkansas has a relatively flat post labor slope (alpha2=.029) and Washington D.C. has a more pronounced descending slope (alpha2=.070). It is apparent from Figure 11a that a2 and alpha2 are correlated. Where a2 (the peak) is higher, alpha2 will necessarily be larger (steeper descending slope), and where a2 is lower, so too is alpha2. In Figure 11b the differences in population distributions are most distinct with regard to the ratio of the population in the middle labor migration years (20-29) versus the early retirement migration years (50-59). In Washington D.C. this ratio is 1.86 and in Arkansas the ratio is 1.35.

Late Years Migration

The c parameter captures the minimum value of the migration profile and this is determined by the level of migration in the late years. Variation in this parameter is clearly linked to the size of the population in the late years, but specifically the relatively youthful late years when there is the highest propensity for migration (ages 65 to 74). Figure 12a shows Florida with the largest c value and Iowa and Utah have similar values for the c parameter, despite the evidence in Figure 12b that shows Iowa with a higher proportion of population in this category than Utah. This may be because the effect of the proportion of the population in this age group on the c parameter may not occur until the proportion gets large enough. For this reason a squared term might be necessary to

define the relationship as curvilinear between the proportion of population in the ages 65-74 and the variation in the c parameter.

V. Estimation Results

The results of the regression models are reported in this section. These are the results of regression models predicting the four early years parameters (Figure 13a). With relatively few variables explain substantial amount of variance in each of these parameters. Each of the effects I explained in the earlier slides were significant in the regression models. DO have some unclear relationships and that's why we call these preliminary results.

The results reported in Figure 13b confirm that the relative sizes of the adult cohorts are significant in their contributions to the middle years parameters. Compare each older group to the one just younger. Alpha2 is positively correlated with a2 (.70). But alpha2 related to the career-retirement ratio. Middle labor years Ages20-29/early retirement ages 50-59

Figure 14 summarizes the regression models in another way. The baseline with the decreasing cohort size has the highest a2 which suggests that if the youngest adult cohort is the largest then career/labor migration will be more peaked. If the adult population has a bulge at 20-24, a2 is smaller, or if the bulge is at 30-34 is larger than 35-39, a2 is smaller. But if the bulge occurs between ages 25-29 then there is an increase in a2. What I gather is that relatively large cohorts at 15-19 or 25-29 will give a higher peak. The early one may be caused by the "going off to college" phenomena or the

continuing education effect. The later one might be caused by the "going off to first professional job" effect.

Figure 1. National Age Pyramids: Mexico, 1970 and Sweden, 1974

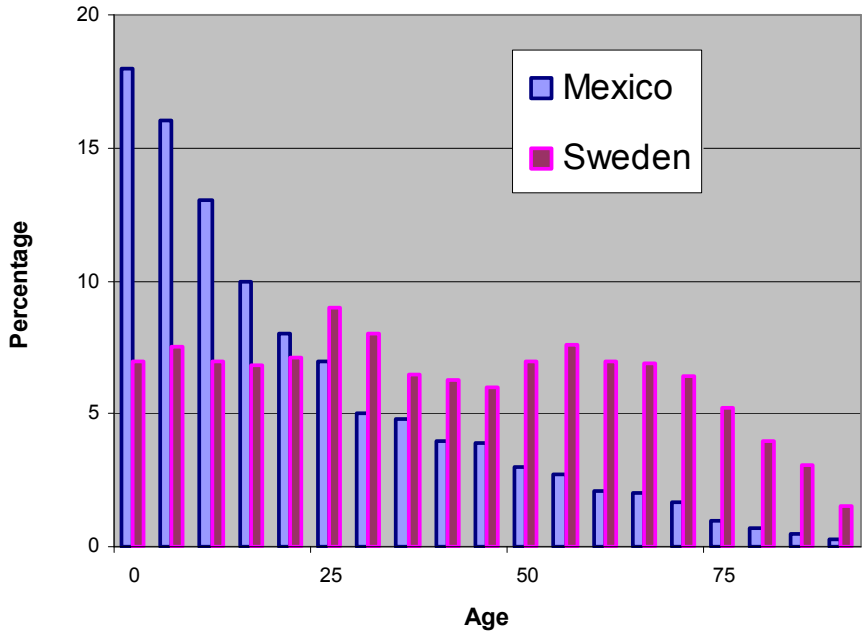


Figure 2. Interregional Migration Profiles: Mexico, 1970 and Sweden, 1974

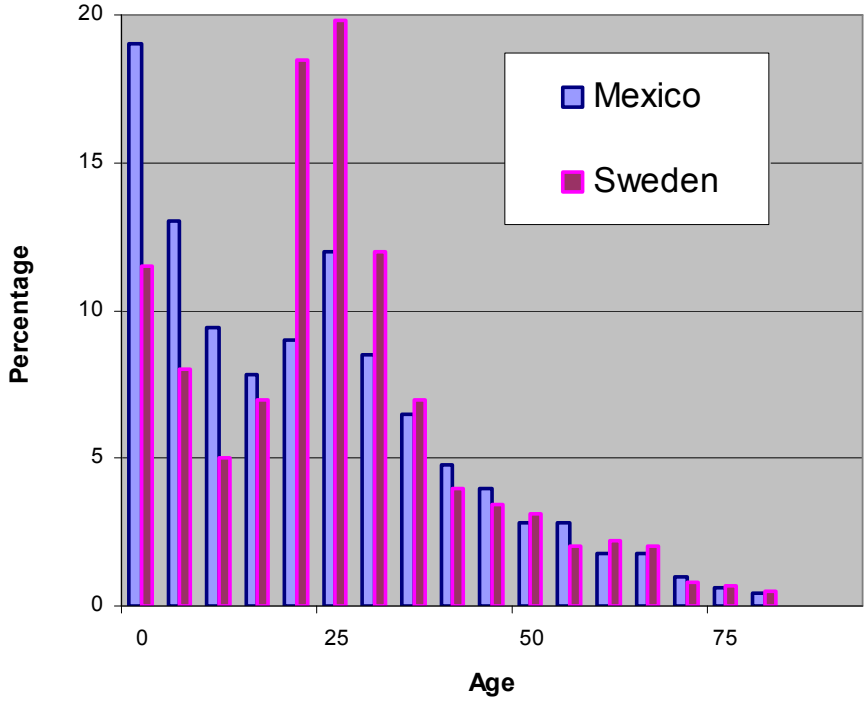


Figure 3. The Rogers-Castro Model Migration Schedule

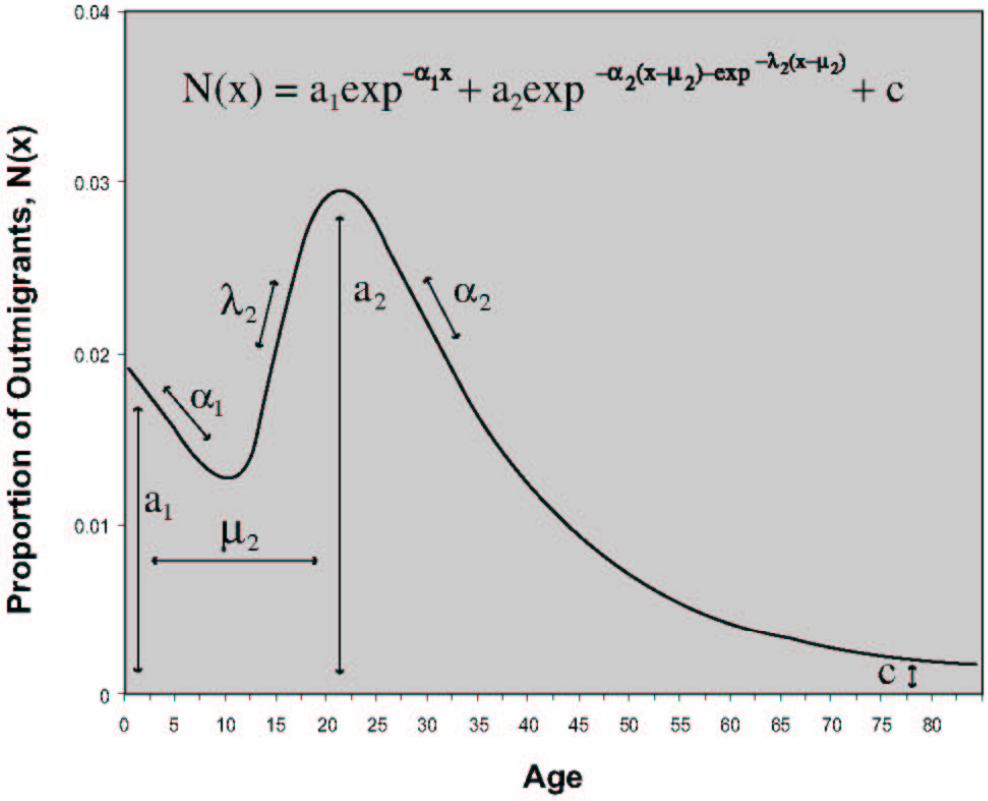


Figure 4a. Maine Migration Profiles (Smoothed Observed vs Model Schedule)

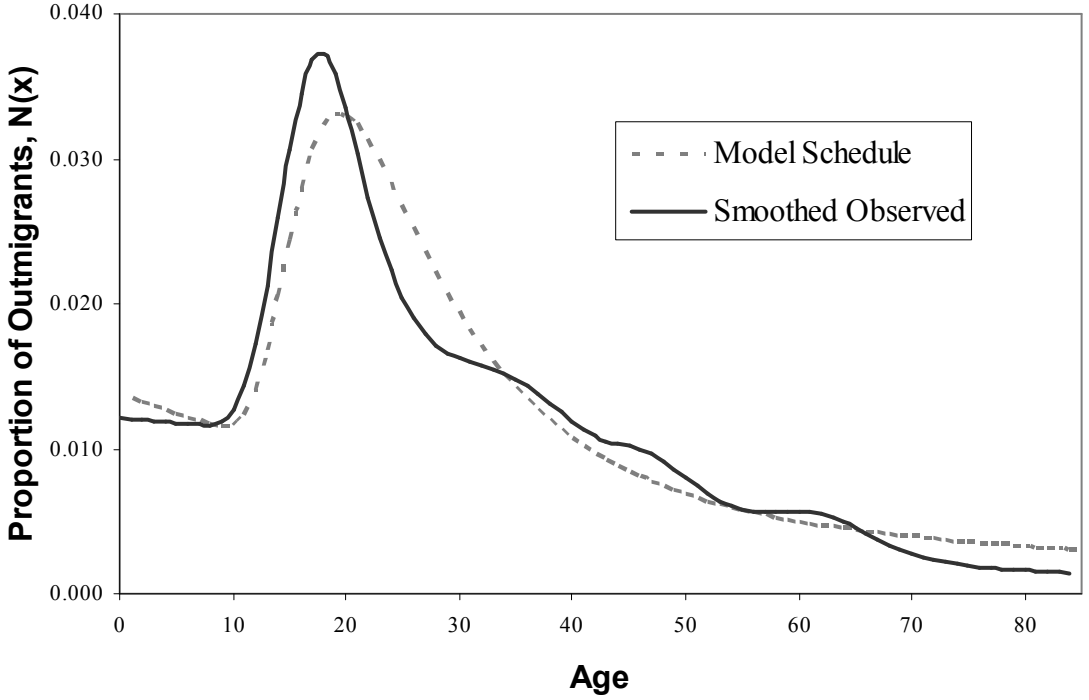


Figure 4b. Pennsylvania Migration Profiles (Smoothed Observed vs Model Schedule)

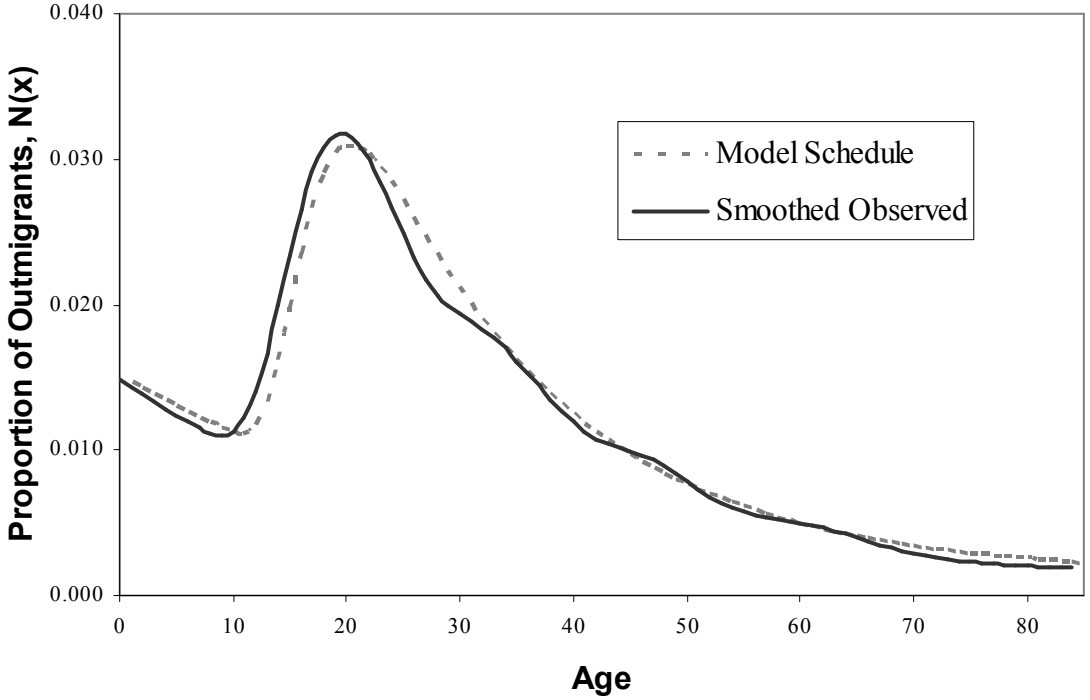


Figure 4c. New York Migration Profiles (Smoothed Observed vs Model Schedule)

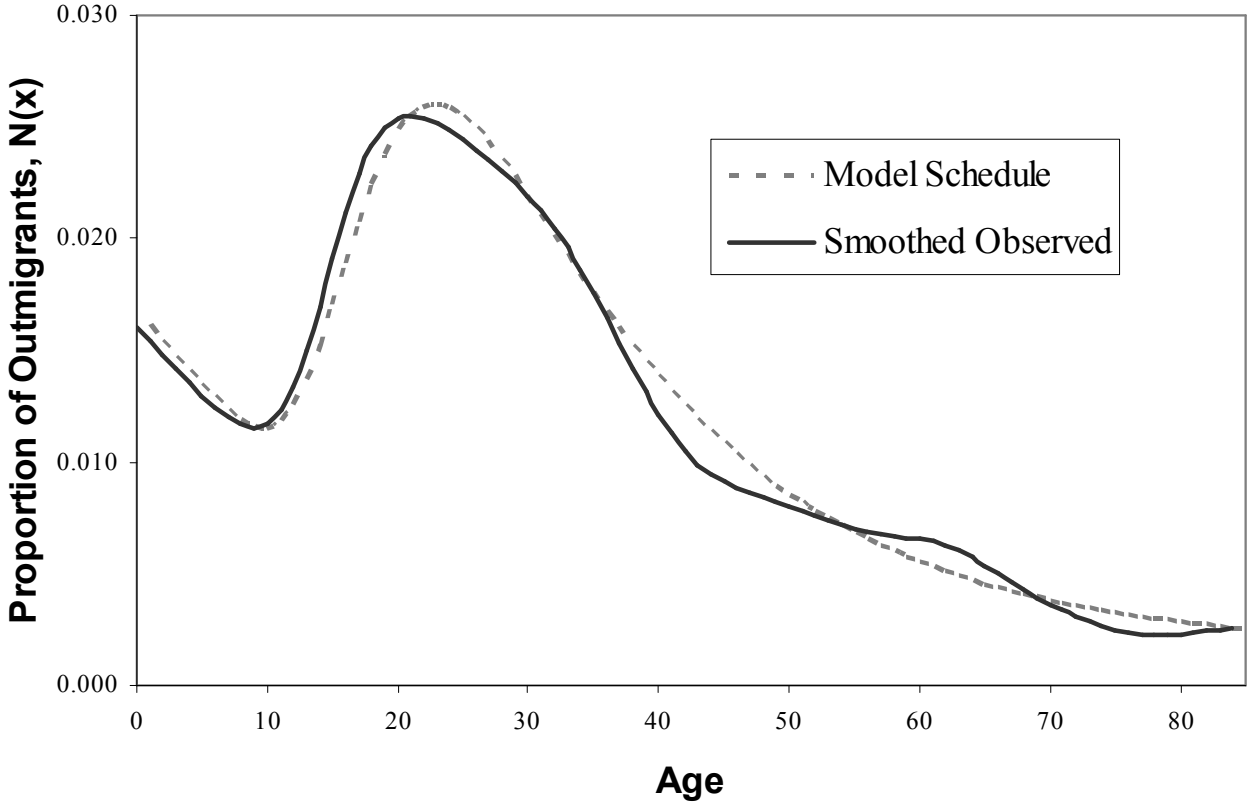


Figure 5. The Rogers-Castro Model Schedules for Outmigration, 51 States

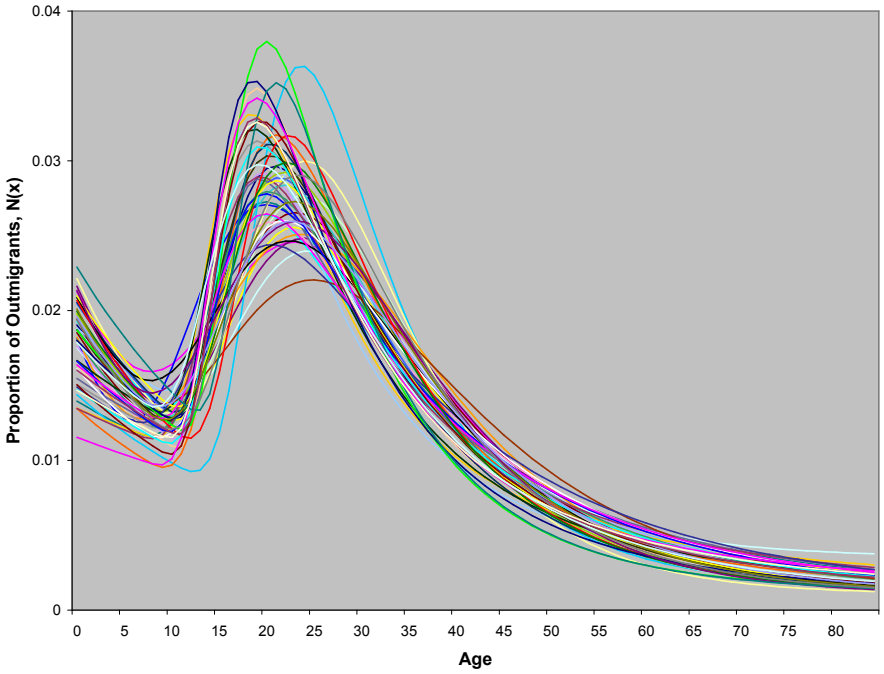


Figure 6. Partitioning the Variance of Model Schedule Parameters

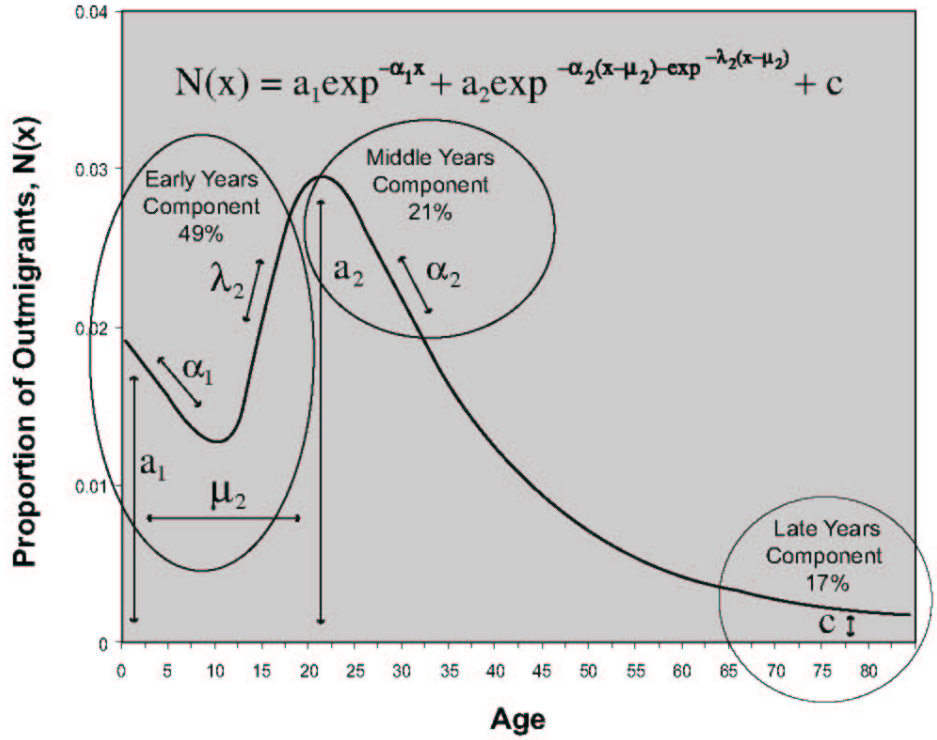


Figure 7a. Variations in the Infant Peak (a1) and the Prelabor Slope (alpha1)

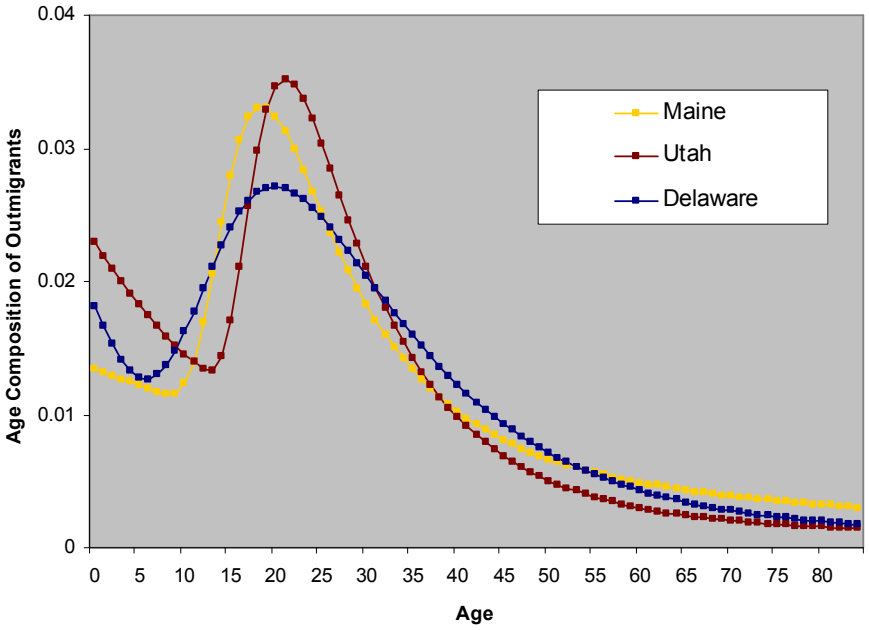


Figure 7b. Population Distributions

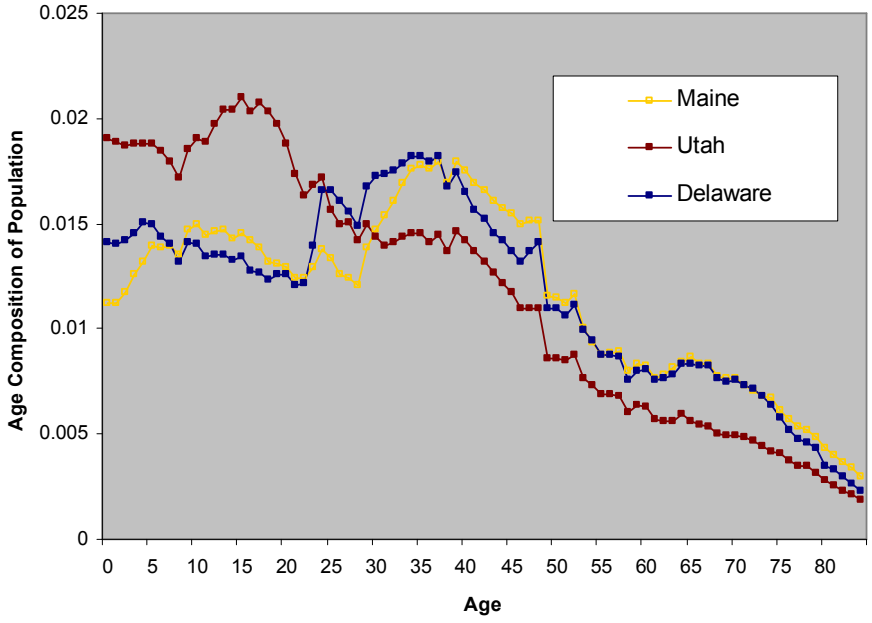


Figure 8a. Variations in the Labor Slope (lambda2)

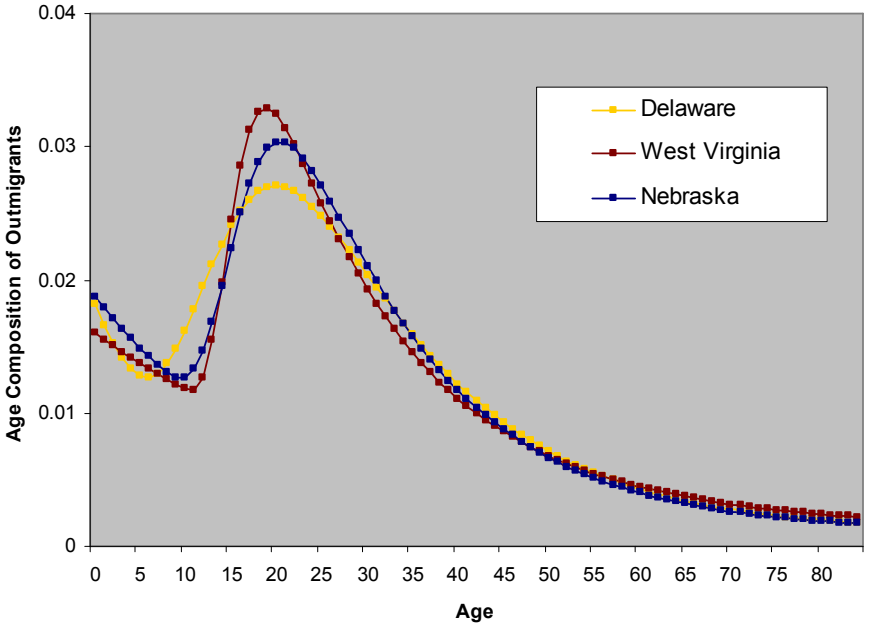


Figure 8b. Population Distributions

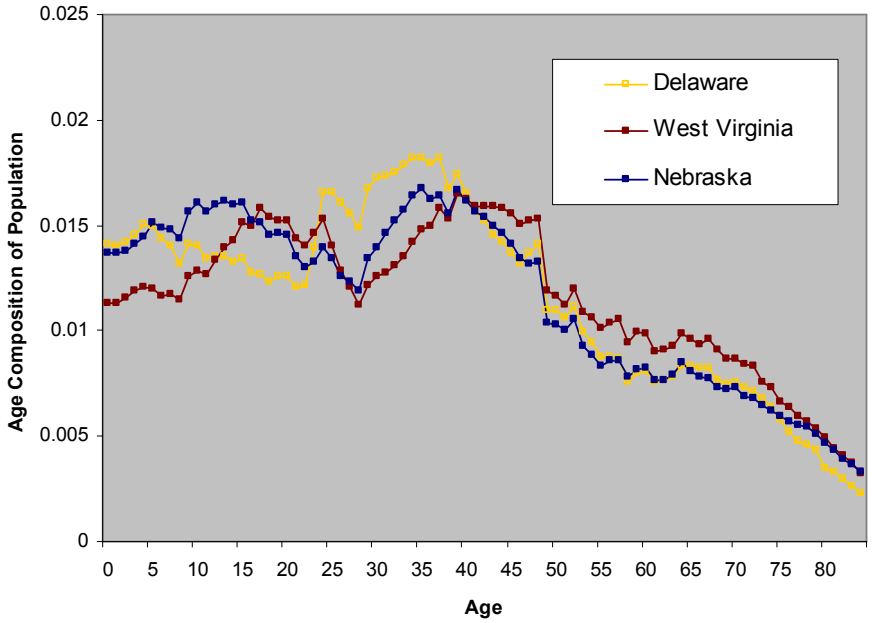


Figure 9a. Variation in the Mean Age of Migration Parameter, Mu2

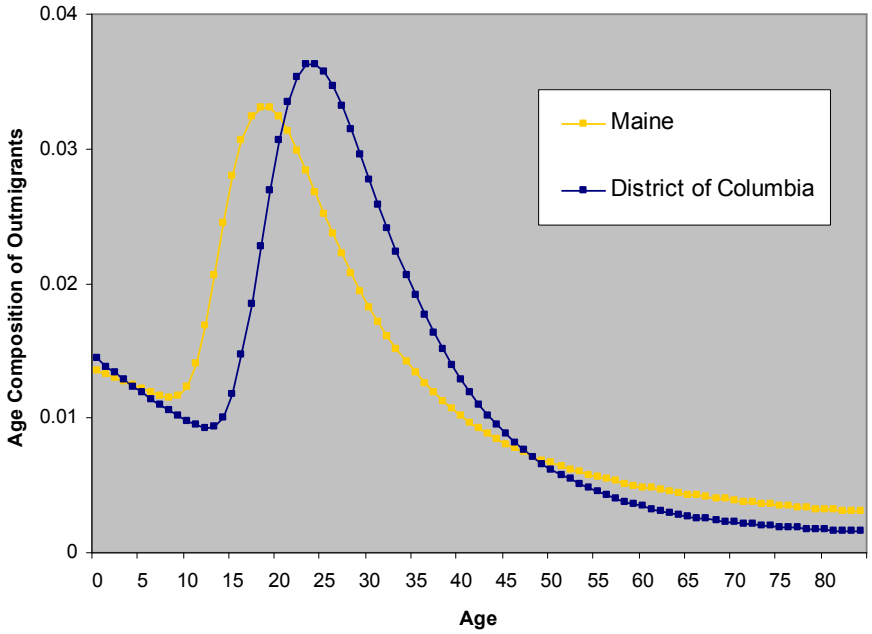


Figure 9b. Population Distributions

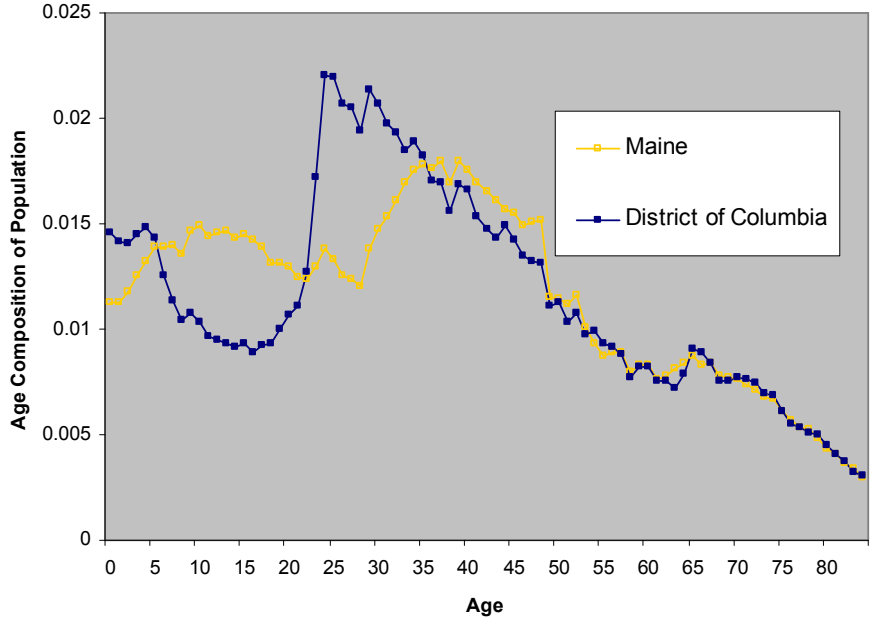


Figure 10a. Variations in the Height of the Career Peak Parameter, a2

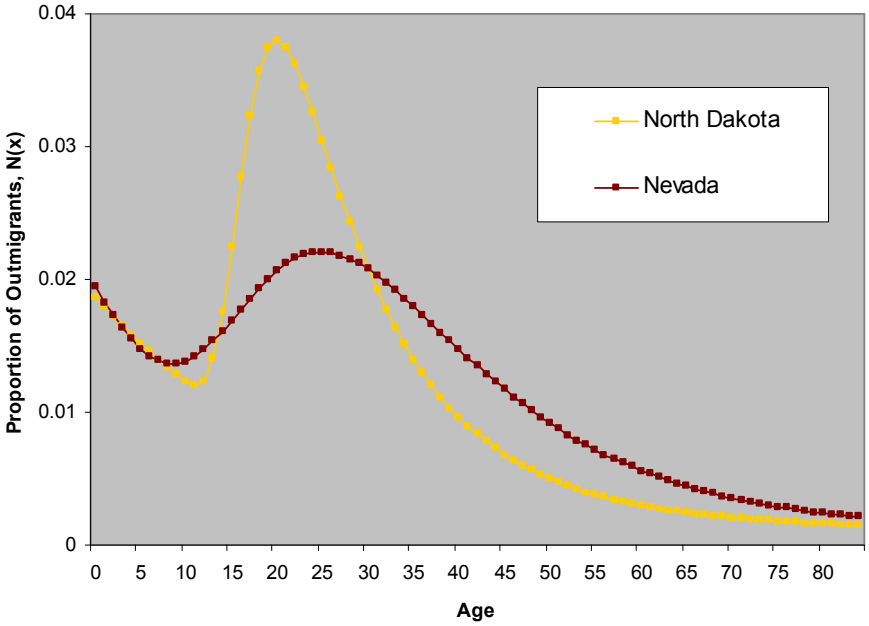


Figure 10b. Population Distributions

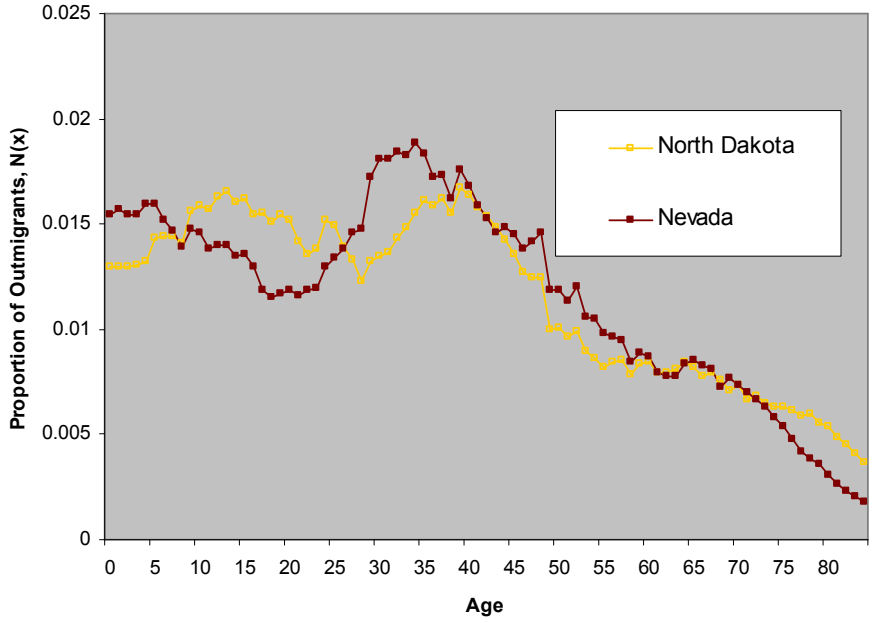


Figure 11a. Variations in the Post Labor Parameter, alpha2

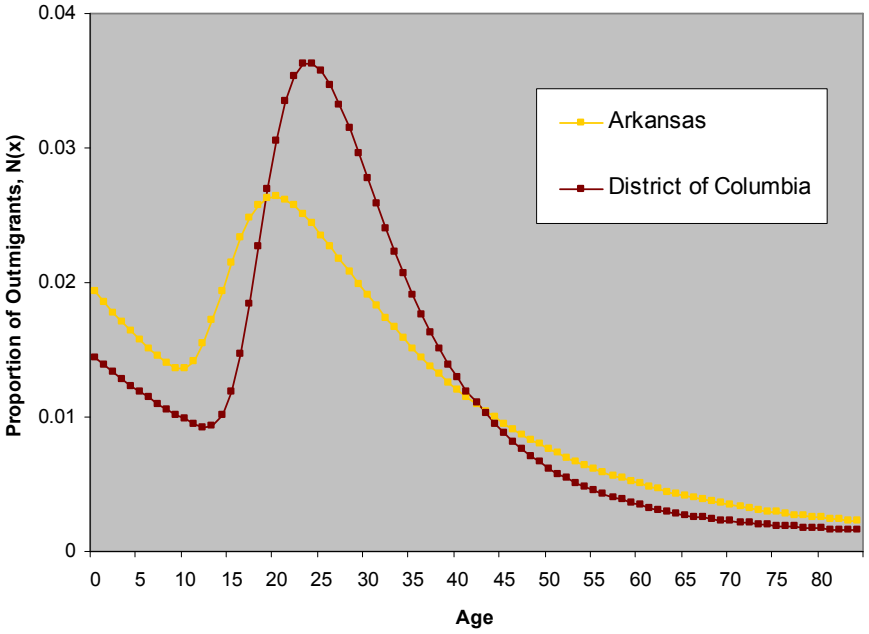


Figure 11b. Population Distributions

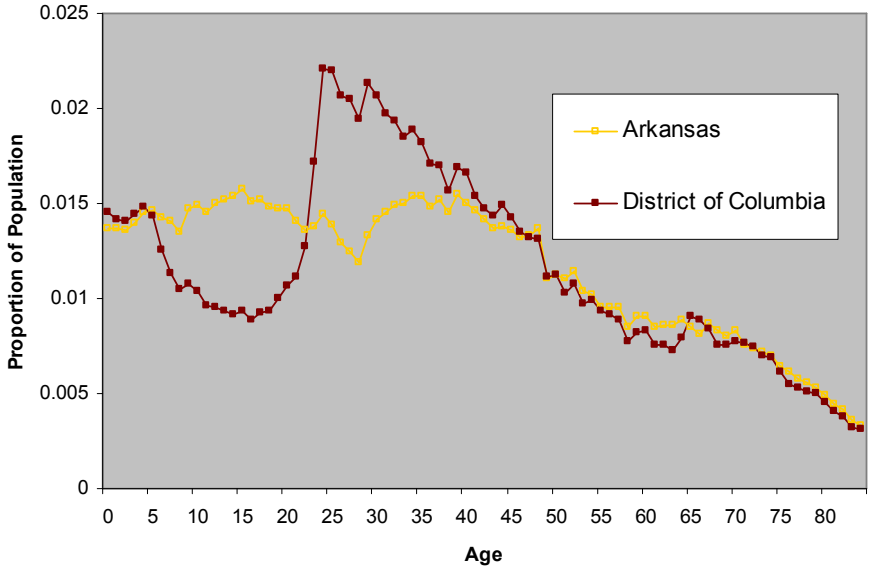


Figure 12a. Variation in Late Year Migration Parameter, c

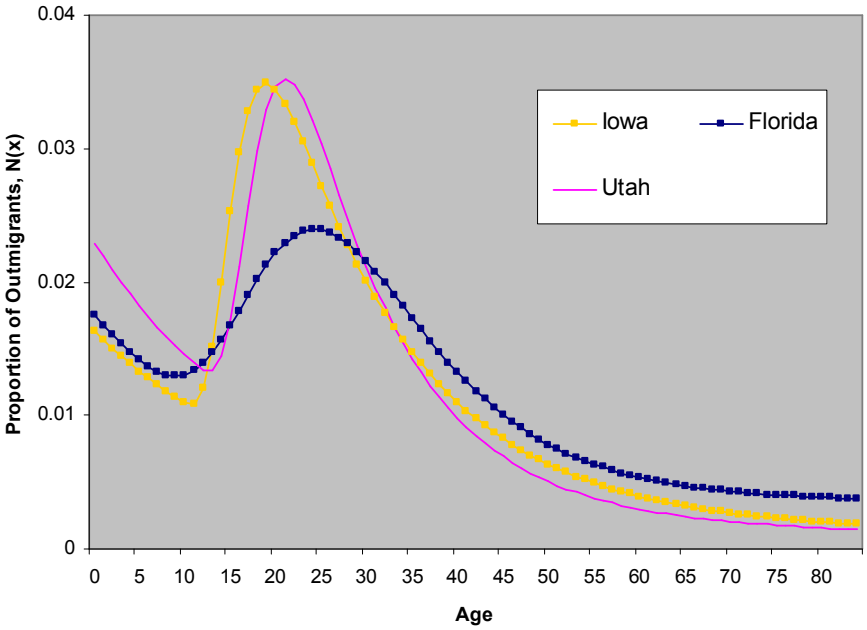


Figure 12b. Population Distributions

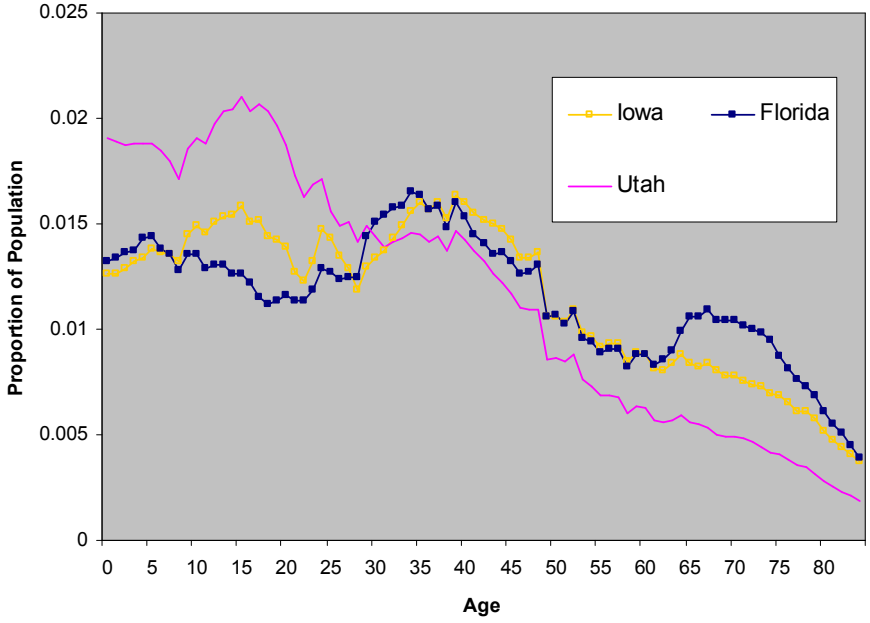
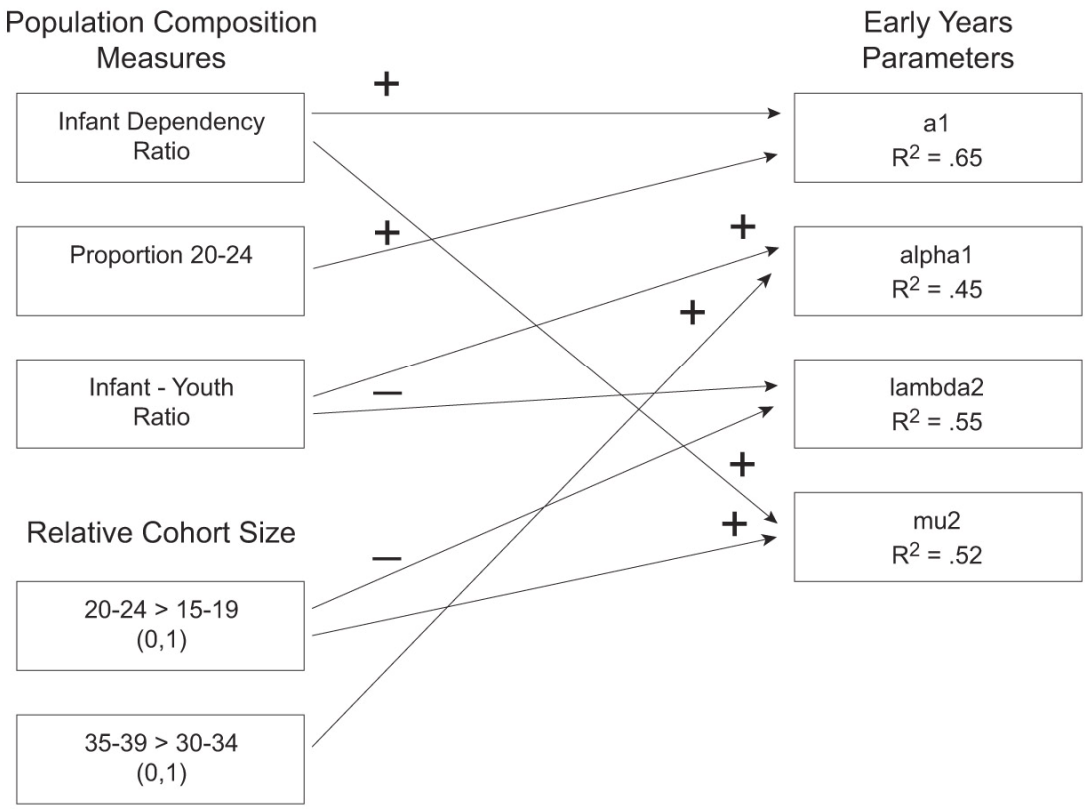
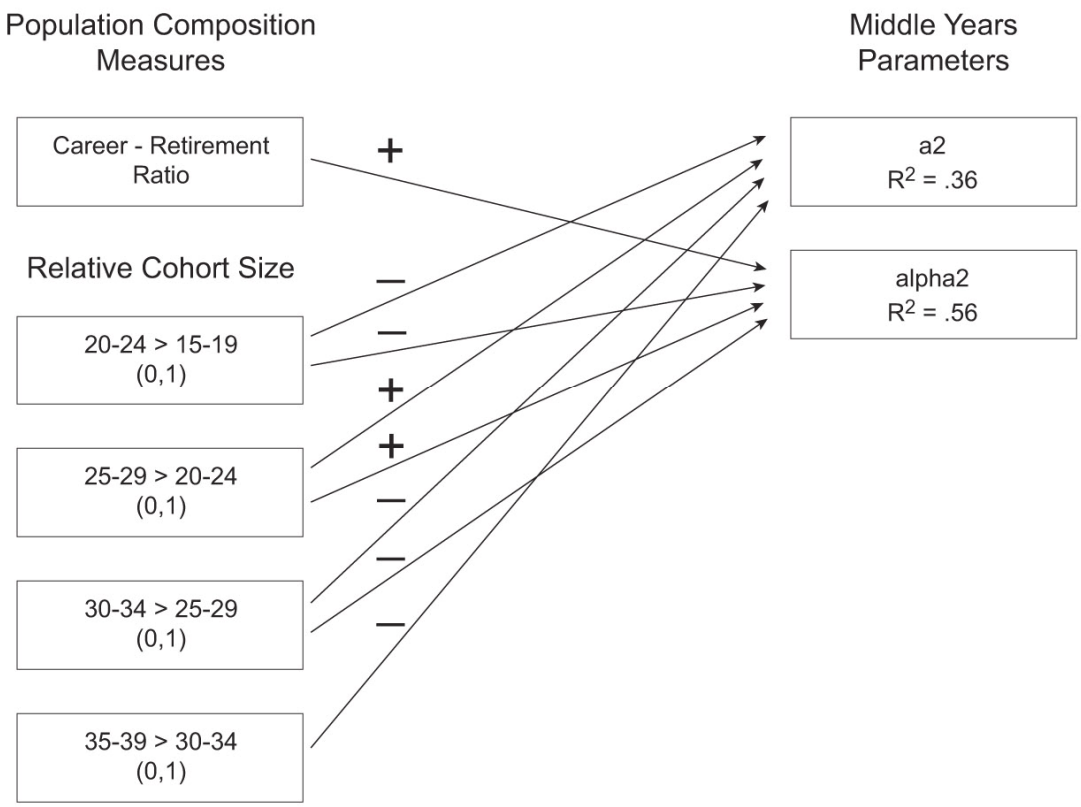


Figure 13a. Predictive Models for Early Years Parameters *



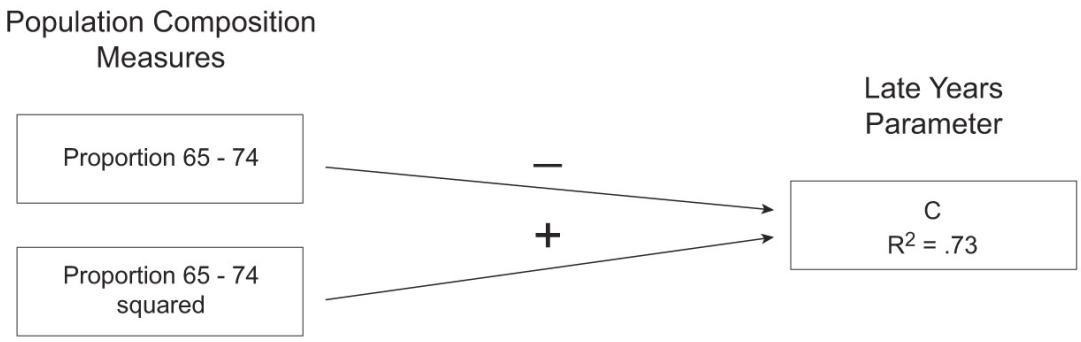
* only statistically significant paths reported

Figure 13b. Predictive Models for Middle Years Parameters *



* only statistically significant paths reported

Figure 13c. Predictive Models for Late Years Parameter *



* only statistically significant paths reported

Figure 14. Adult Cohort Size Effects on a2