Environment and Planning A 2002, volume 34, pages 341 – 359

DOI:10.1068/a33226

Capturing the age and spatial structures of migration

Andrei Rogers, James Raymer
Population Program, Institute of Behavioral Science, University of Colorado, Boulder,
CO 80309-0484, USA; e-mail: andrei.rogers@colorado.edu, raymer@mail.colorado.edu

Frans Willekens
Population Research Center, Faculty of Spatial Sciences, University of Groningen, PO Box 800,
NL-9700 AV Groningen, The Netherlands; e-mail: f.j.willekens@frw.rug.nl

Received 14 December 2000; in revised form 15 August 2001

Abstract. In this paper we model the structures found in the level (generation) and allocation (distribution) components of age-specific and origin – destination-specific migration flows. For the examples, we examine the regional migration patterns in the USA for four periods: 1955 – 60, 1965 – 70, 1975 – 80, and 1985 – 90. The age and migration structures are identified over time by using the logit model for categorical data. Just as model schedules can be used to capture the age patterns of fertility, mortality, and migration rates for use in indirect estimation, so too the models set out in this paper can be used to capture the spatial patterns exhibited by particular sets of age-specific and origin – destination-specific migration proportions. They then also can be used to impose these patterns on inaccurate, incomplete, or otherwise inadequate data.

1 Introduction

In many studies of interregional migration, it is important to have a basic understanding of the underlying age and spatial structures of the observed migration flows. A simple framework for this, based on the logit model for categorical data analysis, is set out in this paper. We use this model for decomposing a set of age-specific and origin – destination-specific migration flows. The goal of this work is to show how the parameters can be used to identify and compare age and spatial structures in migration patterns over time and how these, in turn, can be imposed onto a particular empirical pattern.

Although the methodology and its application to represent migration flows are not new, we believe our particular application of imposing our logit model descriptions of migration spatial structures in the indirect estimation of migration flows has not been suggested before. The reason that such an application will become important is because future US censuses no longer will collect data on internal migration. And the substitution of the American Community Survey's (ACS) continued sample of households will exhibit serious small-sample problems that may need to be resolved by indirect estimation methods.

The notion of structure in age and spatial patterns of migration is relatively well accepted in the field of geography. For example, Tobler (1995) points to both as exhibiting laws of migration. Citing Rogers et al (1978), he argues that the regularity observed in the age structures of migration rates surely warrants designation as a *law* of migration. And then in discussing the spatial structure of migration, he observes that this property is somewhat "sluggish" (that is, stable) in both space and time, presenting a table that "shows the correlation between all six U.S. state-to-state tables for the contiguous United States ... thirty-eight percent of the 1985 – 90 migration table ... can be explained by the 1935 – 40 table, and 52% of it can be explained by the 1975 – 80 table" (Tobler, 1995, pages 336 – 337). This stability is evidence of spatial regularity.

In countries with well-developed data-reporting systems, demographic estimation is based on data collected by censuses and vital registration systems. Demographic estimation in countries with inadequate or inaccurate data-reporting systems often must rely on methods that are 'indirect'. The term 'indirect estimation' is used in demography to describe techniques of estimation that produce estimates of a certain measure on the basis of data that may be only indirectly related to its value, for example, the use of the proportion of children dead among those ever borne by women aged 20 – 24 years to estimate the probability of dying before age 2; or the use of data on the incidence of orphanhood to estimate adult mortality. By identifying the migration age and spatial structures of a particular migration-flow data set, one can then use those structures to indirectly estimate migration flows for situations in which the data are missing or inadequate (United Nations, 1983).

Indirect demographic estimation techniques usually rely on parameterized model schedules—mathematical descriptions of age-specific rates based on patterns observed in various populations other than the one being studied—and select one or more of them on the basis of some incomplete data on the observed population. The justification for such an approach is that age profiles of observed schedules of rate vary within predetermined limits for most human populations. Rates for one age group are highly correlated with those of other age groups, and expressions of such interrelationships form the basis of model schedule construction.

In this paper we propose a set of tools for defining the *spatial* structure of migration that we believe will serve population geographers and demographers in a manner similar to model schedules. We present a unified, general, and parsimonious method for describing the spatial structure of migration. Just as model schedules can be used to capture the *age patterns* of fertility, mortality, and migration rates for use in indirect estimation, so too the models set out in this paper not only capture the *spatial patterns* exhibited by particular sets of origin – destination-specific migration proportions, but also allow these patterns to be imposed in data settings that lack them.

2 Data

Between 1960 and 1990, the population in the USA over the age of 5 years grew by 70.3 million, or from 154.5 million to 224.8 million. Of that growth, 10.3% occurred in the Northeast, 14.3% in the Midwest, 43.1% in the South, and 32.3% in the West. The average annual growth rates for the thirty-year period were 0.0057 for the Northeast, 0.0068 for the Midwest, 0.0165 for the South, and 0.0223 for the West. The regional differences in growth rates were largely a result of internal migration patterns from the Northeast and Midwest to the South and West. The structures of the migration flows that created the particular redistribution of the US population during this time period are the main interest of this paper.

The migration data in this paper represent four time periods: 1955 – 60, 1965 – 70, 1975 – 80, and 1985 – 90. The data were collected from the US Public Use Microdata Series files and describe the number of persons by region of residence at time of census and region of residence five years prior to the census. The regions in the analysis are the Northeast, Midwest, South, and West regions, as defined by the US Census Bureau. The migration data are categorized by region of origin, region of destination, and age, denoted by O, D, and A, respectively. Note that the regions in O and D are denoted by 1, 2, 3, and 4, for Northeast, Midwest, South, and West, respectively, and the seventeen age groups (ages 0 – 4, 5 – 9, ..., 80+) are denoted by the first age in the interval and at the beginning of each five-year migration period. For example, age 20 denotes migrants who were between the ages of 20 and 24 at the beginning of a five-year migration interval (that is, in this case 1955, 1965, 1975, or 1985).

The age-specific migration flow data are expressed in conditional survivorship proportions, conditional because only those who survived to the time of the census could report their migration status five years earlier. Such multiregional conditional survivorship proportions are defined as:

$$\bar{S}_{ij}(x) = \frac{n_{ij}(x)}{n_{i+}(x)}, \qquad j \neq i,$$

where $\bar{S}_{ij}(x)$ denotes the proportion migrating from origin i to destination j at age x (measured at the beginning of the migration interval), $n_{ij}(x)$ denotes the number migrating from origin i to destination j at age x, and $n_{i+}(x)$ denotes the number of persons at age x living in origin i at the beginning of the migration interval. The $+$ sign in $n_{i+}(x)$ denotes summation over all categories of j:

$$n_{i+}(x) = n_{ii}(x) + \sum_{j \neq i} n_{ij}(x) .$$

Furthermore, the origin – destination-specific proportions can be decomposed into two components: the *generation component*, $\bar{S}_i(x)$, and the *distribution component*, $\bar{S}_{j|i}(x)$:

$$\bar{S}_{ij}(x) = \frac{\sum_{j \neq i} n_{ij}(x)}{n_{i+}(x)} \, \frac{n_{ij}(x)}{\sum_{j \neq i} n_{ij}(x)} = \bar{S}_i(x)\bar{S}_{j|i}(x) ,$$

with

$$\sum_j \bar{S}_{j|i}(x) = 1,$$

and where $j\,|\,i$ identifies the value for destination j that is conditional on origin i. When multiplied together, these two components yield $\bar{S}_{ij}(x)$, and where, as before,

$$n_{i+}(x) = n_{ii}(x) + \sum_{j \neq i} n_{ij}(x) .$$

Consider the age-specific and directional-specific out-migration proportions from the Northeast and South during the 1985 – 90 period set out in figures 1(a) and 1(b) (over). These age-specific proportions are disaggregated into generation and distribution components set out in figures 1(c) and 1(d), and 1(e) and 1(f), respectively. The generation components in figures 1(c) and 1(d) are the sums of the migration proportions in figures 1(a) and 1(b) (note that $i \neq j$). Or, in other words, the destination proportions in figures 1(e) and 1(f) multiplied by the generation component in figures 1(c) and 1(d) comprise the multiregional conditional survivorship proportions presented in figures 1(a) and 1(b). For example, consider the out-migrants from the Northeast aged 20 – 24 [figures 1(a), 1(c), and 1(e)]. The multiregional conditional survivorship proportions are $\bar{S}_{11}(20) = 0.9092$, $\bar{S}_{12}(20) = 0.0135$, $\bar{S}_{13}(20) = 0.0543$, and $\bar{S}_{14}(20) = 0.0230$. The generation component is $\bar{S}_1(20) = 0.0135 \; +0.0543 + 0.0230 = 0.0908$. And the distribution components are $\bar{S}_{2|1}(20) = 0.0135/0.0908 = 0.1486$, $\bar{S}_{3|1}(20) = 0.0543/0.0908 = 0.5975$, and $\bar{S}_{4|1}(20) = 0.0230/0.0908 = 0.2539$.

By way of comparison, the corresponding age-specific out-migration proportions from the South are set out in figures 1(b), 1(d), and 1(f). For both the Northeast and South regions, the generation component maintains 'standard' migration age profiles, whereas the distribution components are relatively horizontal. Also, the main difference in the age profiles of the generation component between the Northeast and South is that the Northeast exhibits a retirement peak, whereas the South does not. The distribution components in figures 1(e) tell us that out-migrants from the Northeast overwhelmingly prefer the South, ranging from around 60% for ages 20 – 29 to 85% for ages 60 – 64. Out-migrants

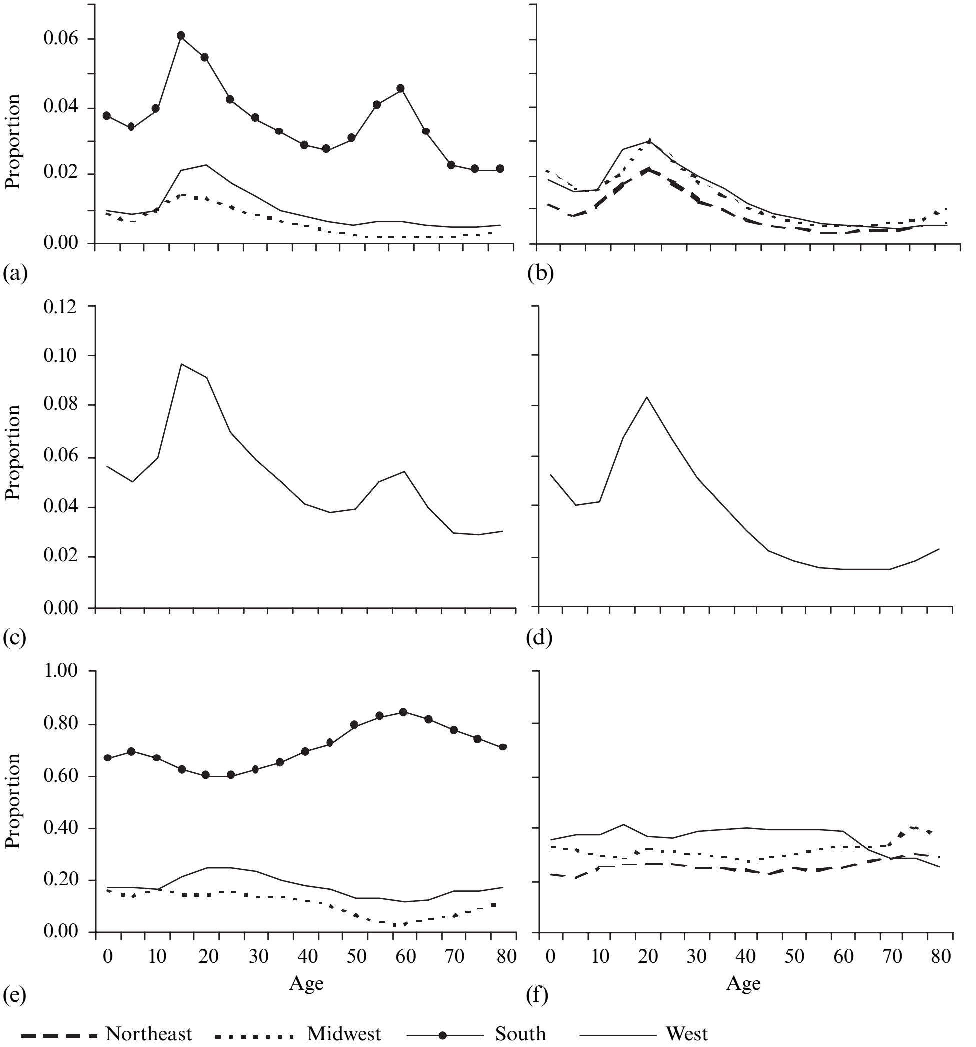

Figure 1. Conditional survivorship proportions of migration from the Northeast and South: 1985 – 90. Multiregional proportions from (a) the Northeast and (b) the South, generation components from (c) the Northeast and (d) the South, and distribution components from (e) the Northeast and (f) the South, respectively.

from the South, on the other hand [figure 1(f)], are not as particular in their destination choices (the West region is slightly preferred for most age groups until the ages 65+).

The regional age-specific generation components for the four time periods are set out in figure 2. The important features to notice in these age profiles are that the Northeast's and Midwest's generation components exhibit retirement peaks, whereas those of the South and West do not. Also, even though there is relative stability over time, there appear to be distinct differences between the first two periods and the last two periods. For example, the generation components for the South region exhibit much higher proportions in the young adult age groups (ages 20 – 29) in the first two periods than in the second two periods. The same seems to be true for the West region, but not so much for the Midwest. Increases in its age-specific out-migration proportions in the second two periods for persons aged 20 and over are evident for the Northeast.

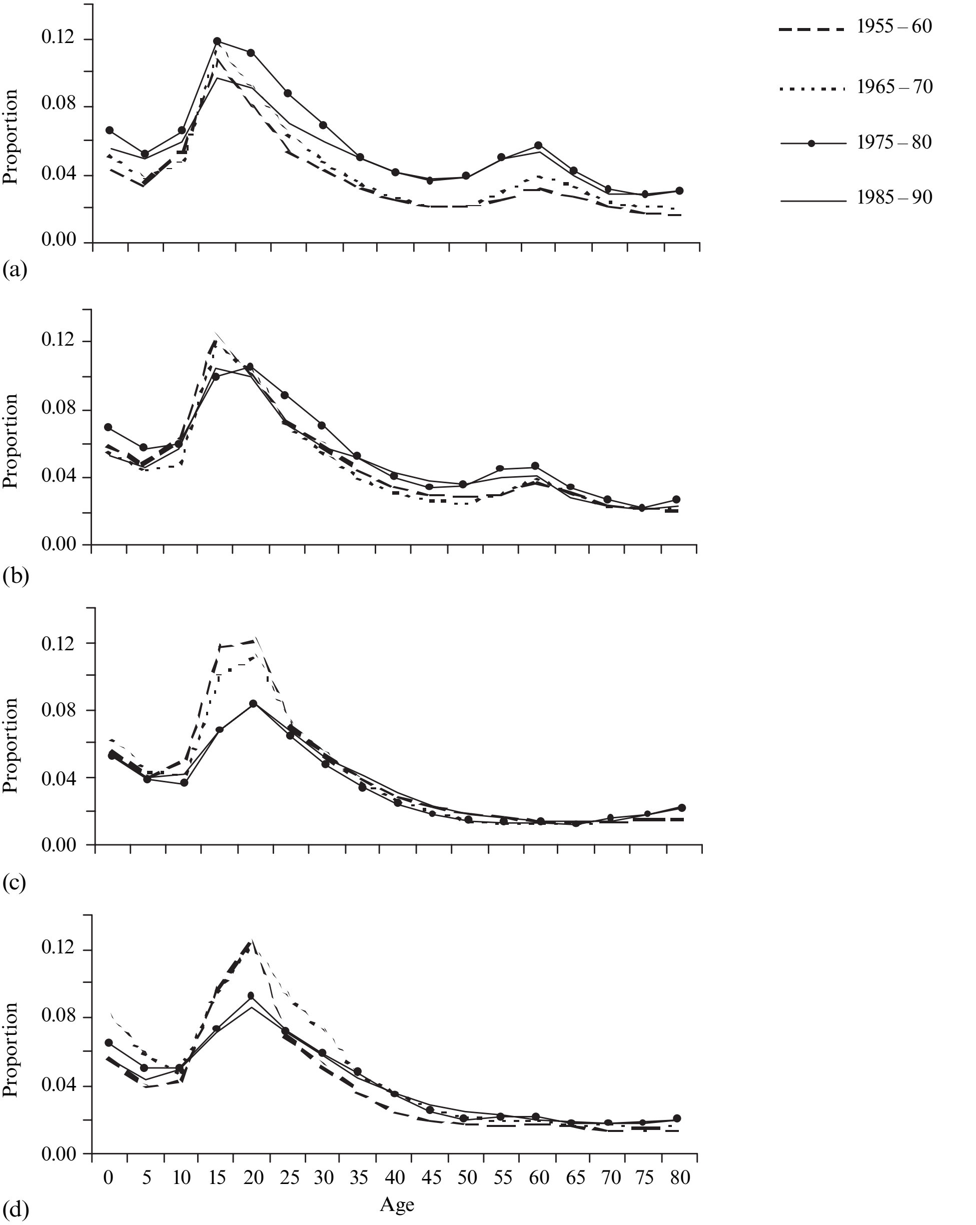

Figure 2. Regional out-migration from (a) the Northeast, (b) the Midwest, (c) the South, and (d) the West: 1955 – 60 to 1985 – 90.

The age-specific distribution components for each region of origin are set out in figure 3 (see over). Apparently, considerable stability was exhibited in the migrant destination proportions for each origin region across the four time periods. The South region was the most attractive destination for out-migrants from the Northeast [figure 3(b)], Midwest (figure 3(e)], and West (figure 3(l)]. These three regions sent, on average, over 50% of their migrants to the South during the study period. Migrants from the Northeast were least likely to go to the Midwest [figure 3(a)], and vice versa

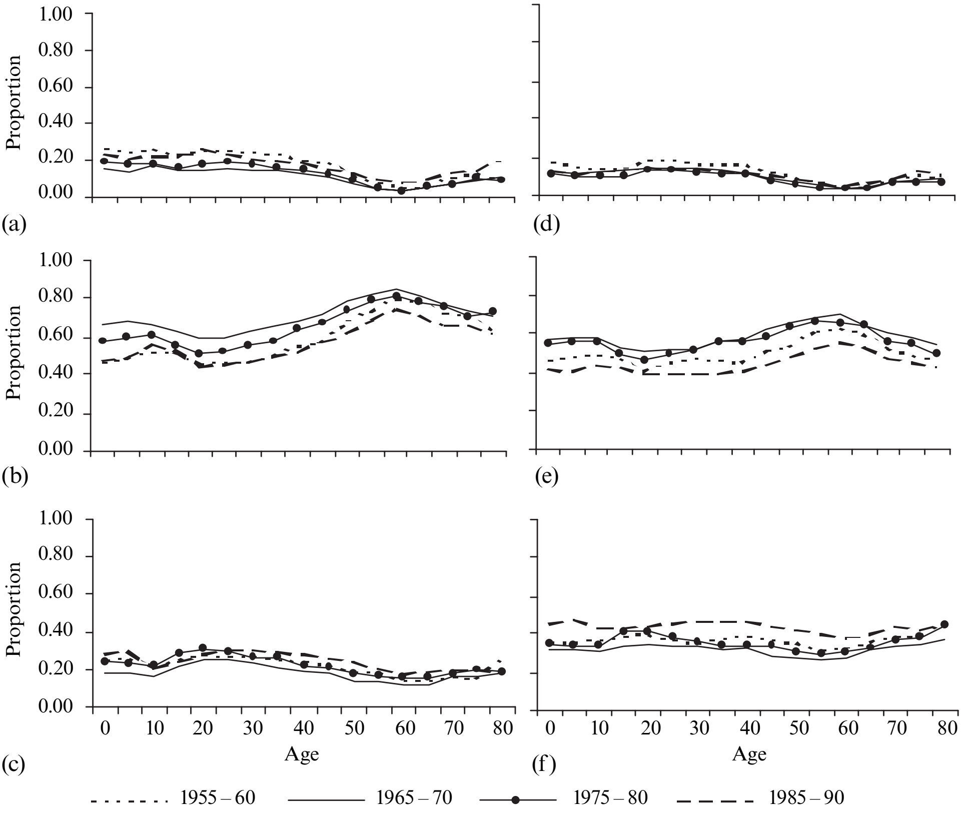

Figure 3. Regional destination proportions 1955 – 60 to 1985 – 90: from the Northeast to (a) the Midwest, (b) the South, (c) the West; from the Midwest to (d) the Northeast, (e) the South, (f) the West; from the South to (g) the Northeast, (h) the Midwest, (i) the West; from the West to (j) the Northeast, (k) the Midwest, (l) the South.

[figure 3(d)]. For the migrants from the South, the most attractive regions were the Midwest and West regions [figures 3(h) and 3(i)].

The age patterns of the destination proportions show that out-migrants from the Northeast and Midwest had higher proportions going to the South in the older age groups (ages 60+) than in the younger age groups [figures 3(b) and 3(e), respectively]. The age-specific proportions from these two regions to the South increased over time, whereas the corresponding proportions to the West decreased [figures 3(c) and 3(f)]. Migrants from the South had relatively flat destination proportions across age groups, except for the South to West flow, in which there was a sharp drop in the proportions after age 60 [figure 3(i)]. These patterns were very stable over time. The most distinguishable features over time of the migrant proportions out of the West were the drop in the proportions of older migrants going to the Midwest [figure 3(k)] and the increase in the proportions going to the South [figure 3(l)].

3 Models

In this section, we set out logit models that capture the structure in the migration data. In general, two types of models are distinguished: saturated and unsaturated models. Saturated models describe the data perfectly, at the cost of a large number of parameters. Unsaturated models are more parsimonious but are not perfect predictors of the data, although they may describe the data with sufficient accuracy. Several types of

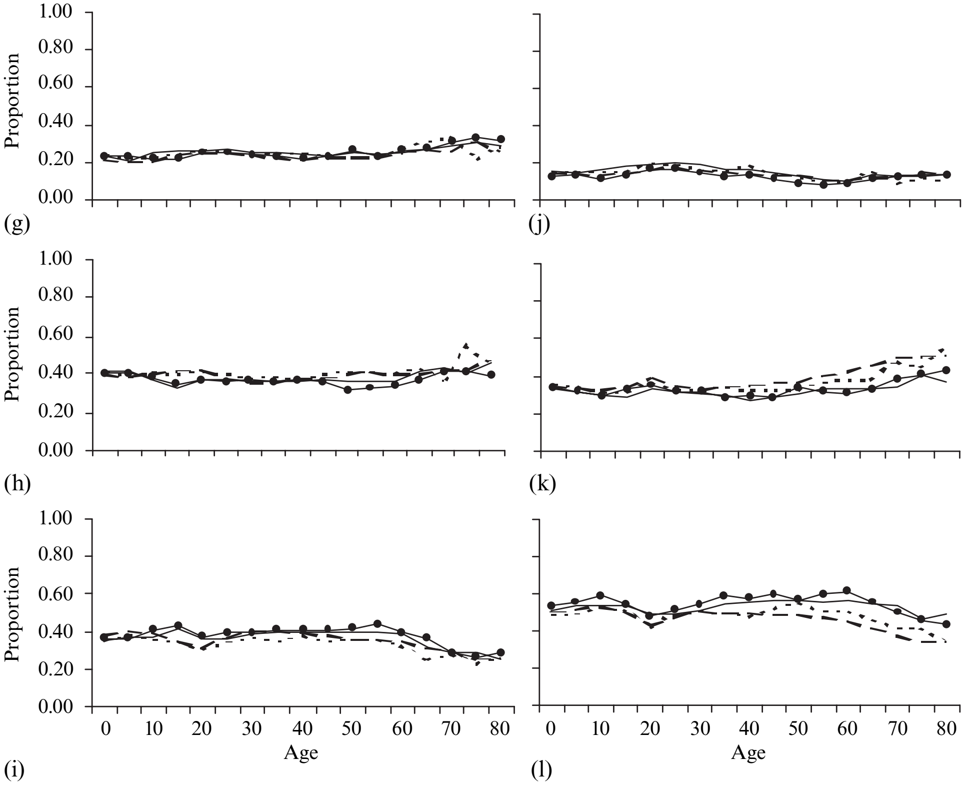

Figure 3 (continued).

models fall within the bounds of the traditional unsaturated and saturated models. For example, as it has been observed that people are much more likely to migrate within regions than between regions, the diagonal elements of migration tables are generally much larger than the off-diagonal elements. Consequently, some models distinguish between diagonal and off-diagonal elements.

In this paper, logit models are presented that describe the age-specific generation and distribution components of migration tables. Their parameters can be used to identify the main components that make up the underlying structure of the age-specific migration proportions. These parameters are presented in their exponentialized (multiplicative) form, which allows them to be interpreted as *odds* and *odds ratios*. The parameters were estimated with the statistical package SPSS, which adopts a procedure that assumes the *last* category to be the reference category. For an introductory discussion on the interpretation of overall effect, main effect, and interaction effect parameters of the logit (or logistic) model, we refer the reader to Agresti (1996), Jaccard (2001), and Long (1997).

Contingency table analysis of migration flow data is often complicated by the tendency of most people to stay in their region of residence during the interval being studied. This produces the relatively much larger values of the diagonal elements in such flow tables— values that tend to dominate the analysis. To deal with this imbalance between diagonal and off-diagonal values, one either controls for the former with, say, structural zeros (Willekens, 1983), or one uses models that separate the diagonal values from the off-diagonal values, for example, with separate level (generation) and allocation (distribution) submodels (Liaw and Rogers, 1999; Plane and Rogerson, 1994; Rogers et al, 2002a).

A saturated multinomial logit model may be used to describe the age-specific multiregional conditional survivorship proportions, $\bar{S}_{ij}(x)$, when the dependent variable is specified as region of destination. The 'predicted' values of this model, denoted by $\hat{\theta}_{ij}(x)$, are equal to the age-specific odds of migrating from origin i to destination j relative to destination k, where k is the reference destination. The predicted odds can then be converted into probabilities, that is

$$\hat{S}_{ij}(x) = \frac{\hat{\theta}_{ij}(x)}{\hat{\theta}_{i1}(x) + \hat{\theta}_{i2}(x) + \ldots + \hat{\theta}_{iN}(x)}$$

with

$$\sum_j \hat{\theta}_{ij}(x) = \frac{1}{\hat{S}_{ik}(x)} .$$

The problem with the model specified above is that the interpretation of parameters to describe migration patterns is complicated by the mixture of stayers (diagonals) and migrants (off-diagonals). A convenient way to get around this problem is to disaggregate the migration flow matrix into two components: a generation component and a distribution component (Plane and Rogerson, 1994, page 204).

3.1 Models for describing the age and spatial structures of migration

The saturated binomial logit model is used in this paper to describe the proportion out-migrating from each origin region (see the generation components set out in figure 2). The saturated multinomial logit model is used to describe the proportion migrating to each destination, given the out-migrants from each origin region (see the distribution components set out in figure 3).

The saturated logit model for the generation component may be written as:

$$\hat{\theta}_i(x) = \frac{\hat{S}_i(x)}{1 - \hat{S}_i(x)} = vv_i^{\mathrm{O}} v^{\mathrm{A}}(x) v_i^{\mathrm{OA}}(x) , \tag{1}$$

for all origins i and ages x, where the v are parameters. This model predicts the odds of being a migrant to being a stayer with respect to origin region i. The corresponding saturated multinomial logit model for the distribution component is specified as:

$$\hat{\theta}_{j|i}(x) = \frac{\hat{S}_{j|i}(x)}{\hat{S}_{k|i}(x)} = v_{j|i} v_{j|i}^{\mathrm{A}}(x) . \tag{2}$$

In this model, $\hat{\theta}_{j|i}(x)$ denotes the predicted odds, given origin i, of migrating at age x to destination j relative to migrating to destination k (reference region).

3.2 Models for imposing the age and spatial structures of migration

If the data are incomplete, auxiliary information may be used to predict migration by age. Let $\bar{S}_i^*(x)$ denote a historical (or hypothetical) generation component and $\theta_i^*(x)$ denote the corresponding odds of being a migrant versus a stayer, that is,

$$\theta_i^*(x) = \frac{\bar{S}_i^*(x)}{1 - \bar{S}_i^*(x)} .$$

The generation component for the current period may be predicted on the basis of, for example, information regarding the *aggregate* total odds of migrants leaving region i and the historical data on the odds of age-specific migration represented by $\theta_i^*(x)$. The model then is given by

$$\hat{\Omega}_i(x) = \frac{\hat{\theta}_i(x)}{\theta_i^*(x)} = \frac{vv_i^{\mathrm{O}} v^{\mathrm{A}^*}(x) v_i^{\mathrm{OA}^*}(x)}{v^* v_i^{\mathrm{O}^*} v^{\mathrm{A}^*}(x) v_i^{\mathrm{OA}^*}(x)} = \frac{vv_i^{\mathrm{O}}}{v^* v_i^{\mathrm{O}^*}} = \omega\omega_i^{\mathrm{O}} , \tag{3}$$

where $\hat{\Omega}_i(x)$ is the ratio between the predicted odds of age-specific out-migration, $\hat{\theta}_i(x)$, and the odds of age-specific out-migration included in the offset, $\theta_i^*(x)$. Note that

$$\hat{\theta}_i(x) = \hat{\Omega}_i(x)\theta_i^*(x), \quad \text{and} \quad \hat{S}_i(x) = \frac{\hat{\theta}_i(x)}{1+\hat{\theta}_i(x)} .$$

The ω denotes the parameters of the logit-with-offset model. The parameters of this model are related to the saturated logit models discussed above [that is, equation (1) with $\omega = v/v^*$ and $\omega_i^{\mathrm{O}} = v_i^{\mathrm{O}}/v_i^{\mathrm{O}^*}$, with the numerator being equal to the saturated logit parameters of the predicted odds of age-specific out-migration, and the denominator being equal to the saturated logit model parameters of the offset. The model in equation (3) is not a saturated model and therefore borrows the age main effect and origin – age interaction effect parameters, $v^{\mathrm{A}}(x)$ and $v_i^{\mathrm{OA}}(x)$, respectively, from the auxiliary (for example, historical) data. To show the relationship between the parameters of the logit model with offset and the saturated logit parameters of the predicted odds of age-specific out-migration along with the corresponding parameters in the offset, notice that both $v^{\mathrm{A}}(x)$ and $v_i^{\mathrm{OA}}(x)$ parameters in the numerator and denominator have asterisks—this implies that they come from the offset and that they are equal to each other (thus, they cancel each other out). The result of the above model is an out-migration pattern that exhibits the level (the *total* regional out-migration proportions) of a current period, but adopts the age profile of the offset (for example, of the historical pattern).

Extending to the discussion above [with regard to equation (3)], we may also include a distribution component of a reference period and hence impose an age profile of conditional destination-specific proportions. In that case, the change in the distribution component is modeled, as follows:

$$\hat{\Omega}_{j|i}(x) = \frac{\hat{\theta}_{j|i}(x)}{\theta^*_{j|i}(x)} = \frac{v_{j|i}v_{j|i}^{\mathrm{A}^*}(x)}{v^*_{j|i}v_{j|i}^{\mathrm{A}^*}(x)} = \frac{v_{j|i}}{v^*_{j|i}} = \omega_{j|i} \,, \qquad (4)$$

where $\hat{\Omega}_{j|i}(x)$ is the ratio between the predicted odds of age-specific migration to destination j relative to k, $\hat{\theta}_{j|i}(x)$, and the corresponding odds included in the offset, $\theta^*_{j|i}(x)$.

Model migration schedules may be used as offsets to model the age-specific dimension of the generation component. The age-specific destination proportions, however, do not exhibit the typical age pattern exhibited by the generation component (see figure 3). Therefore the standard model age profiles are not suited for representing the destination component. Instead, one may select age-specific destination proportions representative of one period and impose them to define the destination components of other periods.

4 Describing age and spatial structures of migration: a comparative analysis

In this section we present a descriptive analysis of the saturated logit parameters of the migration patterns set out in figure 2 and figure 3—that is, of the generation and distribution components, respectively. Our intention is to show how the parameters of the logit model can be used to identify regularities in the migration patterns that are not immediately obvious. Some parameters, for example, might exhibit considerable stability, whereas others may not. This information ultimately can be used to estimate the migration flows based on a minimum amount of information (for example, marginal totals) and on auxiliary data (for example, data from a historical period or from a hypothetical situation).

4.1 Generation component

To demonstrate how the logit parameters may be interpreted, consider the saturated (binomial) logit model for the generation component in equation (1). The parameters were estimated for each of the four time periods: 1955 – 60, 1965 – 70, 1975 – 80, and 1985 – 90. The reference categories for this model, which adopts last-reference-category coding, are the West region and the 80+ age group. Because this is a saturated model, the overall effects, which are equal to 0.0148, 0.0186, 0.0211, and 0.0203 for the four time periods, respectively, denote the odds that a person 80+ years will migrate from the West region. The value of the overall effects over time tell us that the odds of 80+ year-olds out-migrating from the West were highest in the 1975 – 80 period and lowest in the 1955 – 60 period.

The remaining parameters can be divided into the origin and age main effect parameters [v_i^{O} and $v^{\mathrm{A}}(x)$, respectively] and the origin – age interaction effect parameters [$v_i^{\mathrm{OA}}(x)$]. These are set out in figures 4, 5, and 6, respectively. The origin main effect parameters set out in figure 4 represent the ratios of the odds of being a migrant aged 80+ from the Northeast, Midwest, and South to the odds of being a migrant from the West at that age for the time periods 1955 – 60 to 1985 – 90. The odds of being a migrant aged 80+ were substantially higher in the Northeast and Midwest compared with the South and West. Furthermore, relative to the West, the odds of being a migrant aged 80+ increased in the Northeast over time, whereas they decreased in the Midwest.

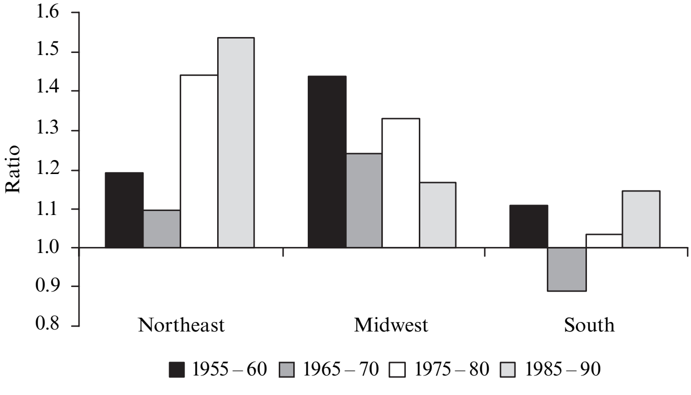

Figure 4. Origin main effect parameters (saturated logit) of the generation component model. Note: model [equation (1)] $= \hat{\theta}_i(x) = vv_i^{\mathrm{O}}v^{\mathrm{A}}(x)v_i^{\mathrm{OA}}(x)$.

The age main effect parameters set out in figure 5 denote ratios of the odds of being a migrant aged x to being a migrant aged 80+ from the West region for the time periods 1955 – 60 to 1985 – 90. Notice that the odds of being a migrant from the West aged 15 – 29 were much higher than those of being a migrant 80+, particularly in the 1955 – 60 and 1965 – 70 periods.

The origin – age interaction effect parameters are set out in figure 6 for the time periods of 1955 – 60 to 1985 – 90. These parameters represent the ratios of the odds of being a migrant from the Northeast, Midwest, and South aged x to the odds of being a migrant from the West aged x. Notice that these parameters have remained relatively stable over time and that the interaction parameters associated with the Northeast and Midwest exhibit peaks in the retirement age groups (ages 55 – 69).

In summary, the parameters of the saturated model tell us (1) that the level of being a migrant aged 80+ has increased over time (from the overall effects), (2) that a person is more likely to be a migrant aged 80+ from the Northeast and Midwest (from the

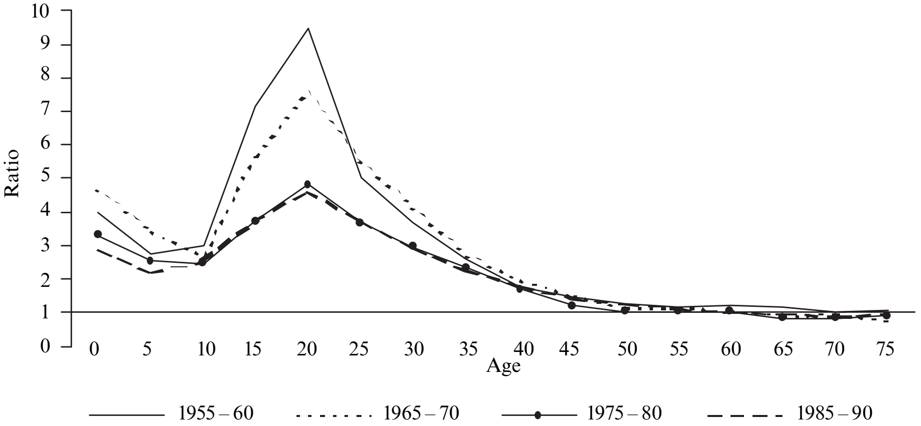

Figure 5. Age main effect parameters (saturated logit) of the generation component model. Note: model [equation (1)] is $\hat{\theta}_i(x) = vv_i^{\mathrm{O}} v^{\mathrm{A}}(x) v_i^{\mathrm{OA}}(x)$.

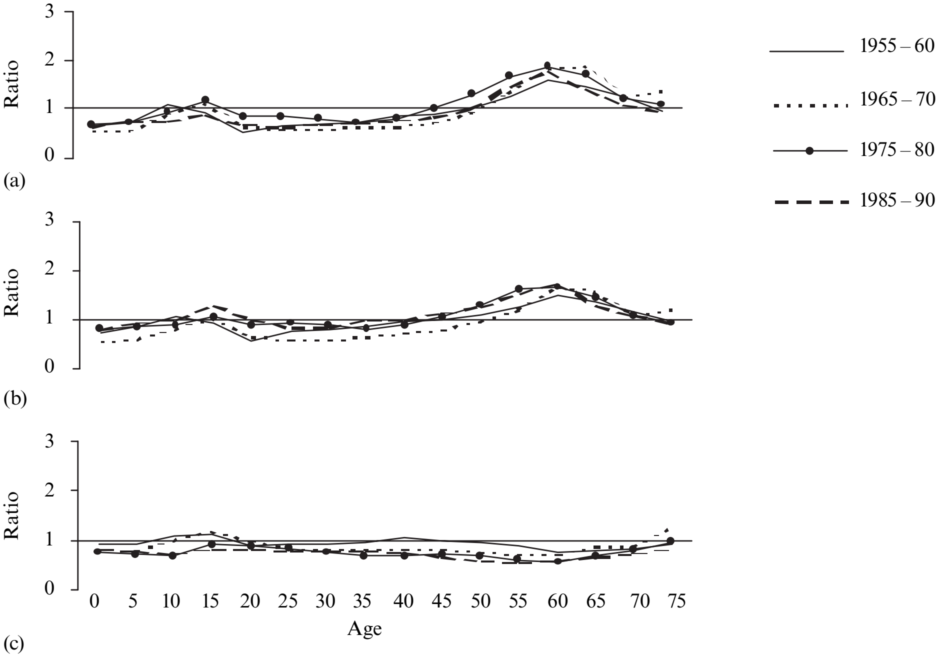

Figure 6. Origin – age interaction effect parameters (saturated logit) of the generation component model: (a) Northeast – age, (b) Midwest – age, (c) South – age. Note: model [equation (1)] is $\hat{\theta}_i(x) = vv_i^{\mathrm{O}} v^{\mathrm{A}}(x) v_i^{\mathrm{OA}}(x)$.

origin main effect parameters in figure 4), (3) that a person has higher odds of being a migrant in the young adult age groups than in the older age groups, which have decreased over time (from the age main effect parameters in figure 5), and (4) that the odds of being a migrant in the retirement age groups are higher in the Northeast and Midwest than in the South and West (from the origin – age interaction effect parameters in figure 6). Most importantly, however, is the stability over time exhibited by the origin – age interaction effect parameters. We can take advantage of this stability to model migration patterns by using more parsimonious models.

4.2 Distribution component

The distribution component can be modeled, once the generation component has been defined. In this subsection we model the distribution component, by age, separately for the four time periods. Because the dependent variable is destination, which has four categories (that is, Northeast, Midwest, South, and West), we use the multinomial logit model. Also, the distribution component is modeled separately for each origin region. This results in some inconsistency in the reference regions. The reference region for the Northeast, Midwest, and South regions is the West, whereas the reference region for the West region is the South. The advantage, however, is that the parameters are more readily interpretable, in the sense that migrants are compared with migrants.

Recall the saturated multinomial logit model set out in equation (2). The parameters are set out in figure 7. The age main effect parameters denoted by $v^{\mathrm{A}}_{j|i}(x)$ represent the ratio of the odds of persons age x years migrating to destination j relative to destination k to the odds of persons aged 80+ years migrating to destination j relative to destination k, given origin i. These parameters are set out in figure 8 (see over).

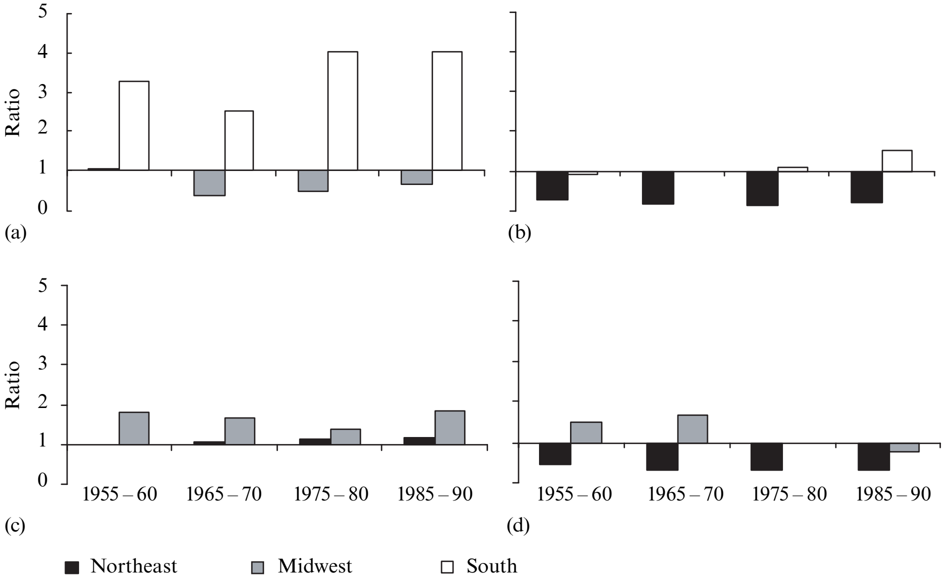

Figure 7. Overall effect parameters (saturated multinomial logit) of the distribution component model. Origin region: (a) Northeast, (b) Midwest, (c) South, (d) West. Note: model [equation (2)] is $\hat{\theta}_{j|i}(x) = v_{j|i} v^{\mathrm{A}}_{j|i}(x)$. The reference region for Northeast, Midwest, and South destinations is the West, whereas the reference region for the West destination is the South.

The overall effect parameters $(v_{j|i})$ in figure 8 (see page 354) provide us with information about the destination choices of 80+ year-old migrants. These migrants from the Northeast are on average 3.5 times as likely to move to the South than to the West, which is their second destination preference. They are least likely to move to the Midwest. The corresponding migrants from the Midwest have been, for the most part, indifferent between the South and West, both of which have much higher destination preferences than does the Midwest. The 80+ year-old migrants from the South, however, are more likely to go to the Midwest and Northeast than the West. Finally, the 80+ year-old migrants from the West have historically chosen the Midwest as their top choice, but more recently, it has been the South.

The age effect parameters $[v^{\mathrm{A}}_{j|i}(x)]$ in figure 8 represent, for each region of origin, ratios of the odds of migrating to destination j relative to moving to the reference destination k at age x to the odds of migrating to destination j relative to moving to the reference destination k at age 80+. In other words, the parameters provide information on how much age matters with regard to destination choice. For example, migrants from the Midwest to the South during, say, the 1975 – 80 period [figure 8(d)] have the greatest odds of migration, relative to the ages 80+, between the ages 50 and 69. The lowest odds (1.0171) of migrating from the Midwest to the South relative to ages 80+ occur between the ages 20 and 24. All age groups 0 – 79 are more likely to migrate to the South than those in the age group 80+. Given this interpretation, we find much variation in the migration flows over time between the Northeast and the Midwest (and vice versa) compared with the other origin – destination migration flows [figures 8(a) and 8(c)]. The South – Northeast, South – Midwest, West – Midwest flows show considerable stability over time.

In this section, we have presented a method for describing migration patterns by focusing on the underlying origin, destination, and age structures and then comparing them over time. This method allows one to identify regularities in the origin, destination, and age effects and the interaction between them in a straightforward and consistent manner. Similar attempts have been made with the related log – linear model (for example, Alonso, 1986; Mueser, 1989; Rogers et al, 2002b; van Imhoff et al, 1997; Willekens, 1983). Although we compared only origin, destination, and age patterns, this method can easily be used to compare any number of additional migration characteristics, for example, period, sex, race, ethnicity, birthplace, and so on.

5 Imposing age and spatial structures of migration: an illustration

For this section, we use the 1955 – 60 generation and distribution components as offsets in the logit model to predict the generation and distribution components in the three subsequent periods. The purpose of this example is to show how a particular structure can be imposed to 'predict' migration patterns of a subsequent period given limited information. Our examples include information only on the marginal totals of a table. The interactions of age with origin and destination are taken from the offset. We show how the parameters of the logit model with the offset are interpreted, and how accurately the 1965 – 70, 1975 – 80, and 1985 – 90 migration flows are predicted by drawing on the spatial structure exhibited by the 1955 – 60 migration flows.

5.1 The generation component

Consider the following unsaturated form of the logit model for the generation component presented in equation (1):

$$\hat{\theta}_i(x) = vv_i^{\mathrm{O}} . \qquad (5)$$

This model predicts the same odds of being a migrant for all age groups. The only data given are the marginal totals, that is, the total numbers of migrants or nonmigrants. Were age main effects added to the model,

$$\hat{\theta}_i(x) \; = \; vv_i^{\mathrm{O}} v^{\mathrm{A}}(x) \, , \qquad (6)$$

the result would produce predicted values with a single 'average' age profile that differed in level according to the total regional differences, which, for example, would predict a retirement peak too small for the Northeast and Midwest and ones too big for the South and West regions. The first model clearly misses the age patterns and is not a good model. The second is better and, with an average age profile, predicts reasonably well (with an R^2 value of 0.94). The problem with the model in equation (6), however, is that it does not distinguish between the different migration age patterns between the Northeast and Midwest as origins and the South and West as origins.

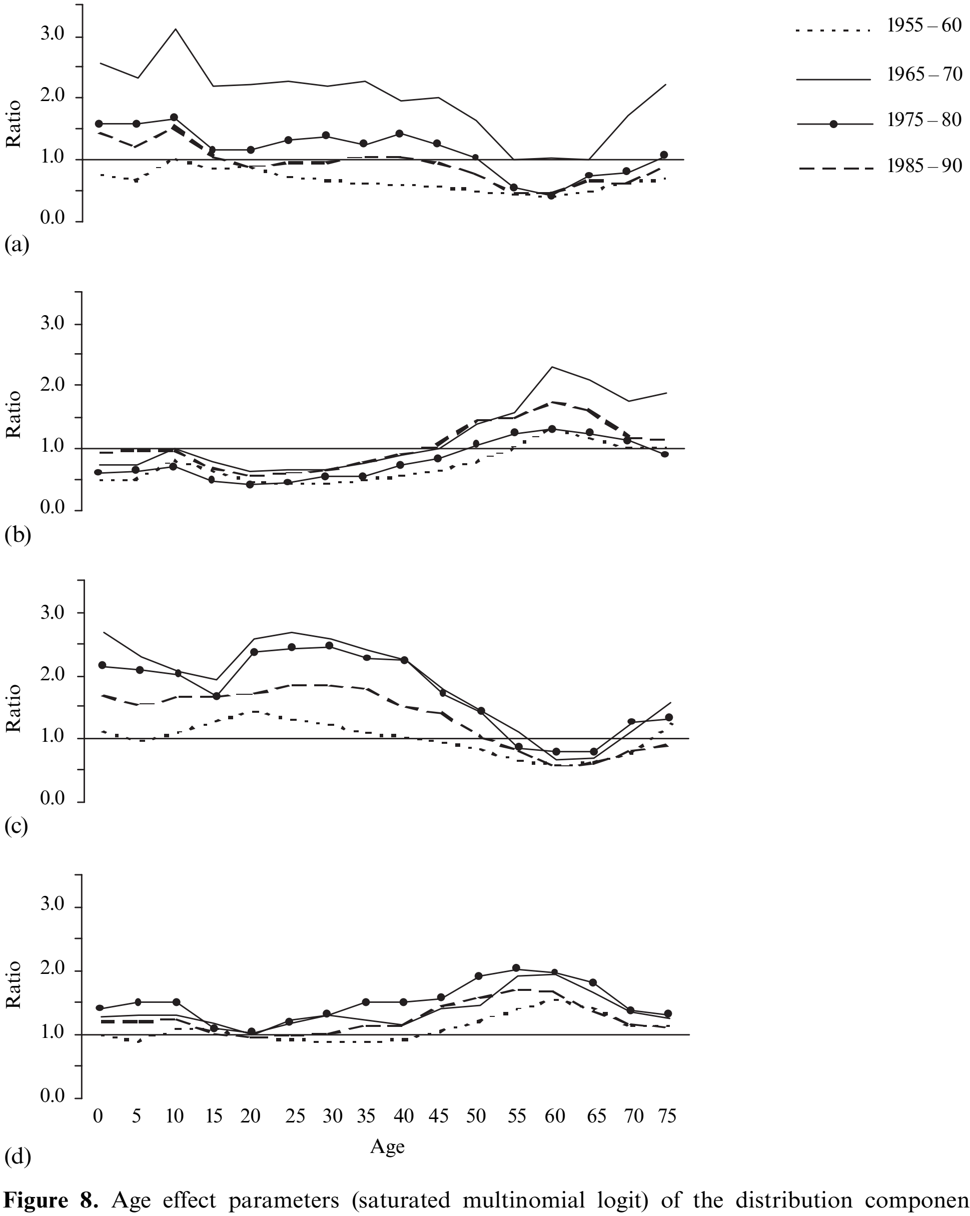

Figure 8. Age effect parameters (saturated multinomial logit) of the distribution component model, for origin–destination flows: (a) Northeast to Midwest, (b) Northeast to South, (c) Midwest to Northeast, (d) Midwest to South, (e) South to Northeast, (f) South to Midwest, (g) West to Northeast, (h) West to Midwest. Model [equation (2)] is $\hat{\theta}_{j|i}(x) = v_{j|i} v_{j|i}^{\mathrm{A}}(x)$. The reference region for Northeast, Midwest, and South destinations is the West, whereas the reference region for the West destination is the South.

An alternative would be to borrow a structure from a previous migration period and fit a logit-with-offset model. Such a model conforms to the same properties as the models above, namely, the proportional totals exhibited in the marginal totals would remain fixed, but would produce age profiles that distinguish between the origins that contain retirement peaks (for example, Northeast and Midwest) and those that do not (for example, South and West).

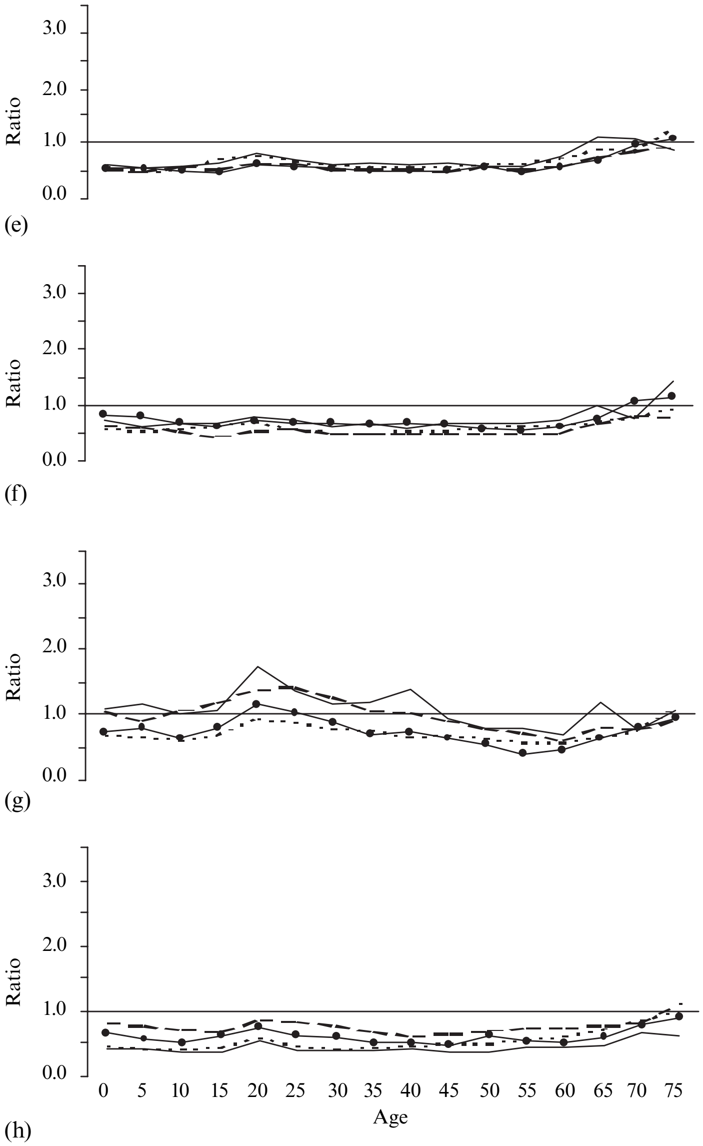

Figure 8b (continued).

The two logit models above [equations (5) and (6)] were applied to predict the 1965–70, 1975–80, and 1985–90 migration proportions with the 1955–60 migration data as an offset. The parameters of these models are set out in table 1 (see over). The first model matches the model set out in equation (3), termed the *region only model*. The second model includes age main effect parameters. This model is termed the *region and age model*, and is specified as:

$$\hat{\Omega}_i(x) = \omega\omega_i^{\mathrm{O}}\omega^{\mathrm{A}}(x)\,.$$

To see the link between the logit model with an offset and the logit models, refer to section 3.2. In the above model, the origin–age interaction effects are taken from the offset (or the 1955–60 period), whereas in the model in equation (5), both the age main

Table 1. Parameters of the binary logit model with offset: predicted 1965–70, 1975–80, and 1985–90 generation components with 1955–60 generation components as offsets.

	Region only model			Region and age model		
	1965–70	1975–80	1985–90	1965–70	1975–80	1985–90
(a) *Parameters*						
Intercept	0.9895	0.9710	1.0254	1.3356	1.3489	1.3581
Region						
Northeast	1.4706	1.3384	1.2320	0.8932	1.4532	1.3188
Midwest	1.0707	1.0076	0.9497	0.7579	1.0598	0.9897
South	0.7611	0.8490	0.7954	0.7648	0.7660	0.8549
Age (years)						
0				1.0301	0.8434	0.7343
5				1.0209	0.8744	0.8039
10				0.7735	0.6783	0.6943
15				0.8615	0.5482	0.5256
20				0.8969	0.6586	0.6003
25				*0.9938*	0.8376	0.7550
30				0.9693	0.8374	0.7813
35				0.9177	0.8421	0.8814
40				0.9275	0.8542	0.9503
45				0.8714	0.8103	0.9252
50				0.8172	0.8471	0.9504
55				0.9366	0.9654	*0.9950*
60				0.9829	0.8784	0.8628
65				0.9471	0.7756	0.7707
70				0.9245	0.8201	0.7829
75				0.9387	0.8234	0.8648
(b) *Goodness of fit* (R^2)						
Northeast	0.9811	0.9318	0.9239	0.9928	0.9810	0.9740
Midwest	0.9816	0.9015	0.9589	0.9838	0.9893	0.9586
South	0.9772	0.9146	0.8917	0.9955	0.9890	0.9760
West	0.9288	0.8979	0.9101	0.9624	0.9587	0.9652
Total	0.9568	0.9161	0.9113	0.9742	0.9743	0.9553

Note: The goodness-of-fit measure (R^2 = coefficient of determination) compares the predicted odds of being a migrant with the observed odds of being a migrant by region and for all flows combined ('total'). The parameter values in italics are not significantly different from unity at the 0.05 level.

effect parameters and the origin–age interaction effect parameters are taken from the offset (that is, the 1955–60 data set).

The parameters of the region-only model are interpreted as the ratio between the predicted parameters and the parameters in the offset. So, for example, the overall effect parameter for the 1965–70 period equals 0.9895. This value is interpreted as the ratio between the predicted overall effect and the overall effect in the offset. Referring to section 4.1, we know that the overall effect of the saturated model denotes the odds of an 80+ year-old being a migrant from the West. We also know that this odds for the 1955–60 period is 0.0148, so the predicted odds must be (0.0148)(0.9895) = 0.0147. The region parameters are interpreted similarly. As shown by the goodness-of-fit measures, these models perform relatively well for the 1965–70 period, but not as well for the 1985–90 period. The offset model performs much better when we include age effects in the model.

5.2 The distribution component

The multinomial logit model with offset for the distribution component is specified in equation (4). Its parameters are set out in table 2. As with the region-only model for the generation component, discussed above, this model adjusts the 1955 – 60 distribution age profiles to fit the observed marginal totals of the 1965 – 70, 1975 – 80, and 1985 – 90 periods. The parameters of the model denote the ratio of the predicted odds of migrating to destination j, relative to destination k, to the corresponding odds observed in the offset (that is, in the 1955 – 60 period). For example, consider the two parameters for the Northeast origin during the 1965 – 70 period (table 2). The parameter for the Midwest destination is 1.1467 and implies that the predicted odds of migrating to the Midwest from the Northeast during the 1965 – 70 period are greater than during the 1955 – 60 period. For the Northeast to South flow, the ratio is not as great, but still there was an increase in the predicted odds relative to those in the offset. Note that the same ratios occur for all age groups because there is no age variable included in the model.

Table 2. Parameters of the multinomial logit model with offset: predicted 1965 – 70, 1975 – 80, and 1985 – 90 distribution components with 1955 – 60 distribution components as offsets.

Origin	Destination	1965 – 70	1975 – 80	1985 – 90
Northeast	Midwest	1.1467	0.7661	0.8429
	South	1.0822	1.1909	1.6945
Midwest	Northeast	1.4401	0.9905	1.3260
	South	1.3346	1.5165	1.8033
South	Northeast	1.1300	0.9433	1.0454
	Midwest	1.0780	0.8713	0.9169
West	Northeast	1.0952	0.8238	1.0524
	Midwest	0.9700	0.8174	0.8048
Goodness of fit (R^2)		0.9374	0.9528	0.9444

Note: The goodness-of-fit measure (R^2 = coefficient of determination) compares the age-specific predicted odds of being a migrant to destination j relative to destination k with the observed odds of being a migrant to destination j relative to destination k, given the origin region i. The reference destination region for the Northeast, Midwest, and South origin regions is the West, whereas the reference destination region for the West origin region is the South. All parameter values are significant from unity at the 0.05 level.

The parameter values set out in table 2 show that the 1955 – 60 distribution components predict the patterns of the 1965 – 70, 1975 – 80, and 1985 – 90 periods relatively well, as demonstrated by the high R^2 values. They also provide information on increases or decreases in the migration patterns between those periods. For example, the ratio of the predicted odds of migrating to destination j relative to k to the corresponding observed odds in the 1955 – 60 period (that is, the offset) were greater for all time periods for the Northeast – South, Midwest – Northeast, and Midwest – South flows (with the minor exception of the Midwest – Northeast flow during the 1975 – 80 period). On the other hand, they were lower for the West – Midwest flow. The other migration flows exhibited more varied patterns over time.

The offset in a logit model is a useful tool for improving estimations in situations where data are missing or inadequate. The offset can come from an historical time period, or it can be constructed from other data sources. To account for changes over time, however, additional variables can be included to improve the model further. For example, in the USA, the census occurs every ten years. Population data for the periods in between censuses are collected by a much smaller survey: the Current Population Survey (CPS). Unfortunately, the CPS is not a reliable source of migration data for small spatial scales and narrowly defined age groups. The method of offsets could potentially produce improved estimates of migration flows by combining the CPS data with prior census migration data.

There are three basic structures of migration according to Tobler (1995) that are confirmed in this paper. The first one is that migration flows exhibit distinct origin–destination-specific patterns, which are relatively stable over time. We illustrate this in our simple examples that used the 1955–60 patterns to help predict those of the 1965–70, 1975–80, and 1985–90 periods. Evidently, most of the migration flow information is captured by the 1955–60 period. The second one is that high correlations between in-migration and out-migration flows exist. This is demonstrated, for example, by the Northeast–Midwest and Midwest–Northeast patterns. The third one is that there are strong regularities in age profiles, which is illustrated by the fact that the majority of migrants are in their young adult ages or, for example, that the elderly have higher propensities to migrate from the Northeast and Midwest to the South.

6 Conclusion

We began this paper with a focus on the data and on the models that underlie the subject of this paper: the description and comparative analysis of the age and spatial structures of migration. The data consist of four consecutive census year counts of interregional migration in the United States; the models are logit models. In section 4 they together provide the vehicle for our comparative analysis of migration structure. In section 5 they lead to a demonstration of how particular structures can be imposed on several observed data sets.

The decomposition of observed conditional proportions (or probabilities) of migration into *generation* and *distribution* components has allowed us to examine regularities, trends, and the relative importance of so-called 'effects': main effects and interaction effects. What we have found is that persistent regularities are exhibited by the age profiles of regional out-migration flows (generation) and by the age-specific destination choices made by these out-migrants (distribution). Together these regularities offer the promise that they can profitably be imposed in empirical studies that lack adequate and accurate migration flow data—a situation that is common in the economically less developed countries, and now in the United States, which no longer will collect decennial data on internal migration in its censuses.

Just as age-specific regularities found in fertility and mortality patterns have led to useful methods for indirectly inferring birth and death rates, so too will the age-specific regularities in migration patterns, which we have identified and represented in terms of parameters, lead to useful methods for indirectly inferring interregional migration rates. But that is a topic for another paper.

Acknowledgements. This research is being supported by a grant from the National Science Foundation (BCS-9986203).

References

Agresti A, 1996 *An Introduction to Categorical Data Analysis* (John Wiley, New York)

Alonso W, 1986, "Systematic and log–linear models: from here to there, then to now, and this to that", DP 86-10, Center for Population Studies, Harvard University, Cambridge, MA

Jaccard J, 2001 *Interaction Effects in Logistic Regression* (Sage, Thousand Oaks, CA)

Liaw K-L, Rogers A, 1999, "The neutral migration process, redistributional potential, and Shryock's preference indices" *The Journal of Population Studies* (published by the Population Association of Japan) **25**(12) 3–14

Long J S, 1997 *Regression Models for Categorical and Limited Dependent Variables* (Sage, Thousand Oaks, CA)

Mueser P, 1989, "The spatial structure of migration: an analysis of flows between states in the USA over three decades" *Regional Studies* **23** 185 – 200

Plane D A, Rogerson P A, 1994 *The Geographical Analysis of Population: With Applications to Planning and Business* (John Wiley, New York)

Rogers A, Raquillet R, Castro L, 1978, "Model migration schedules and their applications" *Environment and Planning A* **10** 475 – 502

Rogers A, Willekens F, Raymer J, 2002a, "Modeling interregional migration flows: continuity and change" *Mathematical Population Studies* forthcoming

Rogers A, Willekens F, Little J, Raymer J, 2002b, "Describing migration spatial structure" *Papers in Regional Science* forthcoming

Tobler W, 1995, "Migration: Ravenstein, Thornthwaite, and beyond" *Urban Geography* **16** 327 – 343

United Nations, 1983 *Manual X: Indirect Techniques for Demographic Estimation* Population Studies 81, Department of International Economic and Social Affairs, United Nations, New York

van E Imhoff E, van der Gaag N, van Wissen L, Rees P, 1997, "The selection of internal migration models for European regions" *International Journal of Population Geography* **3** 137 – 159

Willekens F, 1983, "Log – linear modeling of spatial interaction" *Papers of the Regional Science Association* **52** 187 – 205