Andrei Rogers
Jani Little
James Raymer

SPRINGER SERIES ON DEMOGRAPHIC METHODS AND POPULATION ANALYSIS 26

The Indirect Estimation of Migration

Methods for Dealing with Irregular, Inadequate, and Missing Data

Springer

The Indirect Estimation of Migration

THE SPRINGER SERIES ON

DEMOGRAPHIC METHODS AND POPULATION ANALYSIS

Series Editor

KENNETH C. LAND

Duke University

In recent decades, there has been a rapid development of demographic models and methods and an explosive growth in the range of applications of population analysis. This series seeks to provide a publication outlet both for high-quality textual and expository books on modern techniques of demographic analysis and for works that present exemplary applications of such techniques to various aspects of population analysis.

Topics appropriate for the series include:

- General demographic methods
- Techniques of standardization
- Life table models and methods
- Multistate and multiregional life tables, analyses and projections
- Demographic aspects of biostatistics and epidemiology
- Stable population theory and its extensions
- Methods of indirect estimation
- Stochastic population models
- Event history analysis, duration analysis, and hazard regression models
- Demographic projection methods and population forecasts
- Techniques of applied demographic analysis, regional and local population estimates and projections
- Methods of estimation and projection for business and health care applications
- Methods and estimates for unique populations such as schools and students

Volumes in the series are of interest to researchers, professionals, and students in demography, sociology, economics, statistics, geography and regional science, public health and health care management, epidemiology, biostatistics, actuarial science, business, and related fields.

For further volumes:
http://www.springer.com/series/6449

Andrei Rogers · Jani Little · James Raymer

The Indirect Estimation of Migration

Methods for Dealing with Irregular, Inadequate, and Missing Data

Springer

Prof. Andrei Rogers
University of Colorado, Boulder
Inst. Behavioral Science
Population Program
80309-0484 Boulder Colorado
USA
andrei.rogers@colorado.edu

Dr. Jani Little
University of Colorado
Institute of Behavioral Science
Campus Box 484
80309-0484 Boulder Colorado
USA
jani.little@colorado.edu

James Raymer
University of Southampton
School of Social Sciences
SO17 1BJ Southampton
United Kingdom
raymer@soton.ac.uk

ISSN 1389-6784
ISBN 978-90-481-8914-4 e-ISBN 978-90-481-8915-1
DOI 10.1007/978-90-481-8915-1
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010928637

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To our spouses Mary Ann, Zeke, and Ann Kristin

Preface

This book presents the culmination of our collaborative research, going back over 15 years (Rogers & Little, 1994), and for one of us, even longer (Rogers, 1967, 1973). It addresses a difficult, yet necessary, area of demographic research: what to do in data situations characterized by irregular, inadequate, or missing data. A common solution within the demographic community has been what is generally referred to as "indirect estimation". In our work the focus has been on the indirect estimation of migration, and our use of the term "indirect" follows the description given in the 1983 United Nations manual, which defined it as "techniques suited for analysis of incomplete or defective demographic data" (United Nations, 1983, p. 1).

We wrote this book with a goal to make it accessible to a reader familiar with introductory statistical modeling, at the level of regression and categorical data analysis using log – linear models. It is primarily intended to serve as a reference work for demographers, sociologists, geographers, economists, and regional planners. Space and time limitations have led us to omit topics that some may feel should have been included. In defense, we would argue that our major focus has been on the two principal models of structures in the indirect estimation of migration: model schedules of age patterns and log-linear models of spatial patterns. And our application of this focus has been on three principal aspects of indirect estimation: smoothing, imposing (repairing), and inferring data. Those who are somewhat familiar with the generally accepted methods used in the indirect estimation of fertility and mortality should be able to identify possible extensions of such methods to the case of migration, adopting the fundamental directions outlined by us in Chapters 4, 5, and 6.

Several of the results described in this book first appeared as journal articles or chapters in books. We are indebted to the various editors and publishers for permission to draw on important figures, tables, and paragraphs in those earlier publications. Parts of Chapter 2 first appeared in a 1981 Research Report (Rogers & Castro, 1981) authored by Andrei Rogers and Luis Castro, and its subsequent incorporation as Chapter 5 in the 1986 book on "Migration and Settlement: A Comparative Study" edited by Andrei Rogers and Frans Willekens (Rogers & Willekens, 1986). Both were published by the International Institute for Applied Systems Analysis in Laxenburg, Austria. An early version of parts of Chapter 3 first appeared as "Describing Migration Spatial Structure", written by us in collaboration with Frans Willekens (Rogers, Willekens, Little, & Raymer, 2002). Section 5.2.2 includes some of the results first published in Rogers, Willekens, and Raymer (2003), and Section 6.2.2 summarizes some of the results in Little and Rogers (2007), and Section 6.3.2 draws on results reported in Raymer and Rogers (2007). The rest of Chapters 5 and 6, as well as Chapters 1, 4, and 7, were written especially for this book.

We are grateful to the large number of people who have supported and assisted us in our efforts. Professor Frans Willekens, Director of the Netherlands Interdisciplinary Demographic Institute, introduced us to, and collaborated with us on, the application of log-linear models to describe patterns of interregional migrant flows. Luis Castro was invaluable in the early years of work on model migration schedules. Richard Jessor and Jane Menken, respectively, past and current directors of the Institute of Behavioral Science at the University of Colorado in Boulder, supported us as we pursued this somewhat esoteric topic. Several staff members at the Institute were indispensable in assisting us with the production of the final manuscript: Nancy Thorwardson, Rajshree Shrestha, Julie Na, Lindy Shultz, and Melissa McNabb. A number of graduate students carried out much of the computer processing of our data sets, notably, Lisa Jordan, Bryan Jones, Junwei Liu, and Wanran Ma. And, finally, Andrei Rogers, extends sincere thanks to the National Science Foundation for two supporting grants and to the National Institute of Child Health and Human Development for another grant—all three supporting work on the indirect estimation of migration. In the closing two years of our work, a pilot grant from Colorado University Population Center and general support from its Director, Richard Rogers, greatly facilitated the final drive toward the completion of the manuscript.

Boulder, CO — Andrei Rogers
Boulder, CO — Jani Little
Southampton — James Raymer
December 2009

The Indirect Estimation of Migration Website

A special website – www.colorado.edu/ibs/pop/indirect_estimation_of_migration – has been set up to accompany this book. It is freely accessible and is designed as a resource for students and other users. It contains example programs for estimating the model schedule parameters, the log-linear model with offsets, and the ACS sampling errors. This website will continue to evolve documenting developments and updates that are relevant to the data and the methods used in the book.

Contents

Chapter 1
Introduction

1.1 Introduction

In countries with well-developed data reporting systems, demographic estimation is based on data collected by censuses and vital registration systems. In countries with inadequate or inaccurate data reporting systems, demographic estimation must rely on methods that are more "indirect." Such estimation techniques usually adopt model schedules—parameterized functions describing collections of age-specific rates that are based on patterns observed in populations other than the one being studied—selecting one 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 rates 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.

Although indirect estimation techniques have been applied fruitfully in studies of mortality and fertility, *they have not been developed as systematically and formally for the analysis of migration*. For example, a United Nations manual on the subject is very explicit in its non-coverage of migration:

> A further limitation of the *Manual* is that it deals mainly with the estimation of fertility and mortality in developing countries. There are other demographic processes affecting the populations of these countries (migration for example) which are not treated here (United Nations, 1983, p. 1).

Unlike fertility and mortality, which involve single populations, migration links two populations: the population of the origin region and that of the destination region. This greatly complicates its estimation by indirect methods. What this means in practical terms is that a focus on *age patterns* (as in the case in fertility and mortality) is not enough—one also must focus on *spatial patterns*. And this is where the geographer's particular contribution to migration analysis becomes evident (Isard, 1960; Wilson, 1970; Willekens, 1983a). The imposition of observed regularities in both the age and spatial patterns of interregional migration to "repair" unreliable data on territorial mobility holds great promise as a means for developing detailed

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_1,

age- and destination-specific migration flow data from inadequate, partial, and even non-existent information on this most fundamental process underlying population redistribution.

The principal aim of this book is the description of a formal model-based approach for smoothing, repairing, and inferring directional age-specific migration flows. The approach is one that begins with inadequate or incomplete data on migration, and then estimates "improved" migration data by smoothing, repairing, or borrowing information from other geographical areas, time periods, and data sources. With the elimination of the "long form" questionnaire from U.S. decennial censuses and its replacement by a significantly smaller continuous monthly sampling survey, the American Community Survey (ACS), students of territorial mobility often will find it necessary to deal with inadequate or possibly inaccurate "small sample" data on migration by adopting such indirect methods of estimation. Even more serious "small sample" data problems are being encountered by historical demographers trying to analyze redistribution processes reflected in the sample files of the historical censuses available from the Historical Census Project at the University of Minnesota (www.ipums.umn.edu)—censuses that have the added problem introduced by the total absence of any question on migration (other than the so-called "lifetime migration").

This book includes a focus on formal methods for indirectly inferring migration *flows in the absence of migration data*—for example, from counts of birthplace-specific population *stocks*. The latter have in the past been used to infer patterns of mortality and, indeed, of *net* migration. But no one has developed a workable method for using such population stock data to estimate indirectly *directional* (i.e., origin-destination-specific) migration flows. That is one of the goals of the models described in this book.

The approaches adopted in the following chapters will be useful to at least three user communities: (1) population researchers faced with the loss of the detailed migration data formerly contained in the "long form" questionnaire of past U.S. decennial censuses and its replacement by a significantly smaller continuous monthly sampling survey called the American Community Survey (ACS), (2) historical demographers and geographers seeking to identify changing mobility patterns hidden in the increasingly available historical population censuses that lack a migration question, and (3) migration analysts studying mobility patterns in data poor less-developed countries.

1.2 Models

The estimation of migration from aggregate and incomplete data generally has been carried out with a focus on *net* migration, which is approximated by the population change that cannot be attributed to births and deaths. Given data on population sizes at two points in time, and estimates of birth and death rates for the interval defined by these two points, net migration may be approximated by the difference between the observed population at the second point in time and the hypothetical projected population that would have resulted if only natural increase were added to the initial population. Such methods are reviewed in, for example, United Nations (1967) and Bogue, Hinze, and White (1982).

Methods for inferring *gross* (directional) migration streams have a more limited history (Rogers, 1968, 1975). In the early years, methods of indirect estimation were geared to particular missing data problems. Consequently, the methods had an ad-hoc character (as do many methods of indirect estimation in demography). More recently, however, the indirect estimation of migration has relied on the use of models and on the theory of statistical inference to approximate the parameters from available data. Some models describe *age patterns of migration*, while others describe *spatial patterns of migration* (Rogers, 1999).

1.2.1 Modeling Age Patterns of Migration

Recognizing that most human populations experience rates of age-specific fertility and mortality that exhibit remarkably persistent regularities, demographers have found it possible to summarize and codify such regularities by means of mathematical expressions called *model schedules*.

Over the past 30 years, several studies of regularities in age patterns of migration (e.g., Rogers & Castro, 1981, 1986; Rogers & Watkins, 1987; Rogers & Little, 1994; Rogers & Raymer, 1999; Raymer & Rogers, 2008) have demonstrated that a mathematical expression called the *multiexponential function* provides a remarkably good fit to a wide variety of empirical interregional migration schedules. That goodness-of-fit has led a number of demographers to adopt it in various studies of migration all over the world. The multiexponential model migration schedule (which has become known as the Rogers-Castro model migration schedule) has been fitted successfully, for example, to migration flows between local authorities in England (Bates & Bracken, 1982, 1987), Canada's metropolitan and nonmetropolitan areas (Liaw & Nagnur, 1985), and the regions of Japan, Korea, and Thailand (Kawabe, 1990), and South Africa's and Poland's national patterns (Hofmeyr, 1988; Potrykowska, 1986, 1988). Statistics Canada, for example, has adopted this model migration schedule to produce its provincial population projections (George, 1994). Other examples include analyses of interregional migration in Indonesia (Muhidin, 2002) and international migration in Europe (Raymer, 2007).

1.2.2 Modeling Spatial Patterns of Migration

Spatial interaction (e.g., migration) patterns have been modeled by gravity models, entropy models and, more recently, by log-linear models. The relation between the gravity model, entropy maximization and log-linear models is discussed by Willekens (1980, 1982a, 1982b, 1983), Bennett and Haining (1985), and Aufhauser and Fischer (1985), among others. The relation between the entropy method and log-linear modeling was shown by Good (1963). The relation between the iterative proportional fitting (IPF) method and log-linear models was utilized in estimating migration flows from incomplete data by Drewe and Willekens (1980) and Nair (1985), among others. Finally, Willekens and Baydar (1986), Stillwell (1986), Yano (1991), and O'Brien (1992) all adopt the perspective of generalized linear models (GLMs), a perspective that includes the log-linear model.

Log-linear models are statistical devices that are useful for *describing* and *decomposing* patterns underlying matrices of spatial flows. They are not replacements for theoretical explanatory models that purport to account for observed patterns of behavior. But as instruments for identifying observed regularities in such patterns and then introducing them in settings wherein some data are incomplete, they offer great promise. In particular, they allow for the combination of data from different sources, e.g., censuses, surveys, and administrative records. The models used to do this belong to the family of generalized linear models (GLMs), which include the log-linear model, the logit model, the Poisson regression model, and the logistic regression model. The parameters indicate the contribution of the partial data to the predicted migration flow. The parameters are estimated from the data by maximizing the likelihood that the model reproduces the observations.

1.2.3 A Model-Based Approach to Migration Estimation

Modern methods of multiregional population projection (Rogers, 1975, 1985, 1995) require a migration data set that is quite detailed. Such detailed data are not available in some instances and have to be "created" using indirect estimation methods. This was the situation faced by the U.S. Census Bureau, and was the principal motivation for their efforts to create an adequate "synthetic" migration data set on the basis of inadequate data on migration:

> . . . combines annual geographic information on recent migration from tax return data, information on the relationship between 1-year and 5-year migration rates from CPS, and data on interaction between geographical and demographic dimensions contained in the 5-year interstate migration data from the 1980 census (Wetrogan & Long, 1990, p. 36).

In combining the demographic and geographic detail of the decennial census counts, with the timeliness and frequency of the Current Population Survey (CPS) and matched Internal Revenue Service (IRS) tax returns, the Census Bureau followed the outlines of the "3-Face-Problem" first outlined by Willekens, Por, and Raquillet (1981) and Rees and Willekens (1986). Conceptually, this formulation may be viewed as a cube in which the three axes represent demographic detail, geographic detail, and temporal detail, respectively. Such a formulation leads naturally to the notion of a model-based approach to the indirect estimation of migration—one which is aptly captured by Constance Citro, a study director for the Committee on National Statistics of the National Academy of Science, who writes:

> I define "model-based estimation" loosely, as the use of statistical methods to produce "indirect estimates" for an area by combining data from several areas, time periods, or data sources to "borrow strength" and improve precision. (Citro, 1998, p. 40)

This is the model-based approach to indirect estimation of migration that we adopt in this book. Its philosophy echoes the approaches already being used by the Census Bureau and can be formalized by a model-based methodology that focuses on the statistical analysis of data with missing values. In the approach we use methods and techniques for describing migration and its age and spatial patterns that could improve the quality of estimates of migration flows. This objective is sought by adopting (1) model schedules of migration that describe observed regularities in age structures and (2) log-linear (generalized linear) spatial interaction models that describe regularities in spatial structures in ways that could aid the indirect estimation of directional migration flows in settings with irregular, inadequate, or unavailable data. Both sets of models may be used to apply regularities reflected in other data sets to "discipline" the data being studied.

The above two categories of models may be used in at least three different ways, depending on the quality of the available data. If the migration data are generally reliable, but somewhat irregular, then *smoothing* the data may be a sufficient solution. If, on the other hand, the migration data are clearly non-conforming and unreliable, then they need to be repaired—a process that *imposes* regularities found in other, more reliable, data sets on the inadequate data under study. Finally, if migration data are totally missing or unavailable, then methods for indirectly *inferring* their values need to be used.

We use the term "smoothing" to represent the process of limiting the effect of randomness on the age, spatial or temporal patterns of migration caused by natural variation or variation due to insufficient sample size. This may involve (1) fitting *splines* to observed data across age or over time, (2) fitting a curve to an age profile of migration, or (3) removing higher-order interaction effects in a log-linear model for a contingency table of migration flows. We use the term "imposing" to represent the process of borrowing age or spatial patterns of migration from other regions (e.g., when an average age profile of out-migration from a Census Region in the U.S., such as the West, is used to represent the age profile of out-migration from a small state in that region, such as Wyoming) or drawing on an auxiliary migration data source of somewhat comparable measurement of migration. The second option often involves a situation where only marginal information is available for a matrix of flows, and more detailed data are borrowed from a recent census or survey. We use the term "inferring" to represent the process of borrowing age and / or spatial patterns from an auxiliary data set, one that nevertheless can be used to serve as a proxy for the particular migration pattern that we are estimating (e.g., when tax return data are used to infer migration).

1.3 Data

Observations on migration can be of very different types, and they are often incomplete, i.e., some required information is missing. In the presence of incomplete data, we propose a strategy that first identifies the data types of migration.

1.3.1 Observed Data: Data Types

Migration involves a relatively permanent change of residence address, one involving the crossing of an administrative boundary. The definition and the measurement of migration involve both a time dimension and a space dimension. Moreover, it is useful to distinguish *event data*, i.e., data on the event of migration, and *status data*, i.e., data on the place of residence at two (or more) points in time. Event data typically describe the number of events (*migrations*) that occur during a given time interval. Status data typically relate to the numbers of persons, in a given location, who lived in a different location at some prior date (*migrants*), or to those who are expected to live in a different location at a future date. They are often expressed as proportions rather than counts. The distinction between event data and status data yields two broad data types that require different modeling strategies of migration. Since in this book only transition data are analyzed (except for the Swedish data in Chapter 2), we call such data on transitions, migration.

Another distinction is between micro-data and grouped data. Micro-data are data on individuals (or households). They are typically associated with surveys. Grouped data are aggregations of individual data. Within the above broad categorizations we shall generally restrict our attention to status data on migrants and to grouped (transition) data.

Given the above data types, various degrees of incompleteness may be considered. For example, the number of migrants over an observation interval may be recorded, but some information may be missing. Or, information on some attributes (covariates) of migrants may be missing for all persons (e.g., age). Missing attributes are important when they explain individual differences in the risk of experiencing migration. Information on particular variables may be observed for some persons and not observed for others. Attributes may be partially missing, e.g., as a result of non-response. Areal units for which data are required may not be the units for which data are available. Also boundaries may have changed, calling for areal interpolation or extrapolation.

The responses to missing data may use auxiliary data and/or information on comparable populations (e.g., expressed in the form of model schedules). The selection of auxiliary data and of standard schedules are important topics in expositions of how the missing information may be estimated from available data. Generally, one develops a probability model for the complete data and estimates parameters of the model from incomplete data and the auxiliary data. The assumption is that the different observations are manifestations of the same underlying mechanism.

1.3.2 Using Auxiliary Data

The basic strategy in estimating detailed migration levels and patterns is to use as much information as possible on the actual migration patterns to be inferred. Often the information comes in the form of aggregations of the detailed migration patterns to be determined. Aggregation may be over space, age, or time. For instance, detailed interstate migration data may be lacking, whereas data on interregional migration may be available. The age composition of migrants may be available for the state of origin only, but not by state of origin *and* state of destination. To disaggregate the data to the desired level of detail, theoretical distributions or empirical distributions may be applied, or a combination of both. The distributions impose a structure onto the migration patterns which are absent in the primary set. Model schedules are such theoretical distributions. A historical migration table at the desired level may serve as an empirical distribution. The migration patterns of infants or children below the age of 5 may also serve as an empirical distribution. Theoretical and empirical distributions may be combined, e.g., the migration patterns of those under-five may be combined with model migration schedules and primary data in order to determine migration flows by age.

In most applications, the imposed structure captures empirical regularities and therefore represents useful knowledge in the estimation of missing data. If it is plausible to assume that the missing data have a structure that does not deviate from what is generally observed, then a theoretical or a borrowed empirical structure may be applied to improve the estimates. Technically, the theoretical and empirical distributions may be integrated in the model as *offsets* (Rogers et al., 2003). In statistical models, offsets are used to fix parameters or regressions coefficients to given values, while other parameters or coefficients are estimated from the data. Parameters are often fixed because they have known values, or because the available data do not permit their estimation. The latter case applies when data are incomplete and the parameters are "borrowed" from the auxiliary data. Needless to say, the quality of the estimates is dependent on the plausibility of the auxiliary information. Our methods draw on auxiliary information for the estimation of migration from incomplete data.

Three categories of auxiliary data may be considered. The first consists of historical data. Historical data usually come from a previous census. The higher-level spatial and/or age structures (e.g., the higher-order interaction effects exhibited in contingency tables) indicated by the historical data may still be valid even when they are outdated. Higher-level structural changes usually lag behind changes in levels and structural changes at lower levels. The second category consists of contemporary migration data from the same or a comparable population. For instance, surveys and the census provide information on migration that may be combined effectively using a modeling approach. Detailed data on migration among a subpopulation (e.g., those under five years of age) may provide the structure that can be applied to other subpopulations (e.g., people aged 20–24 years). The third category consists of judgmental data (e.g., expert opinions) about migration. Different categories of auxiliary data may be combined to generate the preliminary estimates of migration that encompass a higher-order structure of migration patterns, one which may be imposed onto incomplete primary data in order to obtain more plausible and reliable migration estimates.

1.3.3 The Case of No Migration Data

For data settings of *no* available data on migration, the indirect estimation of migration may be carried out using census data on birthplace-specific population stocks for one or more points in time, particularly for infants. In such a data setting we will first need to obtain an "initial estimate" of the migration regime—an estimate that may be further refined by imposing regularities described by our model migration schedules and our log-linear models.

Embedded in censuses or survey enumerations that generate distributions of persons cross-classified by age and place of current residence is "hidden" information about the migration patterns that helped to shape such distributions. We have developed several different promising perspectives that we believe yield adequate crude estimates of age- and origin-destination-specific migration flows from such data on spatially distributed survivors of region-specific birth cohorts. For example, it is possible to use a model to link observed propensities of *infant migration*, inferred from birthplace-specific population stocks, to the associated propensities of all other age groups (Rogers & Jordan, 2004). Such a procedure is tested in Chapter 6. It is also possible to turn to a procedure that decomposes residually estimated *net migration* flows into the underlying gross in- and out-migration flows (Rogers & Liu, 2005), or to adopt a procedure that borrows observed past regularities in the relative intensities of *secondary* (return and onward) versus *primary* migration streams to indirectly estimate migration flows (Rogers & Raymer, 2005). Both of these procedures need further refinement, and hence are not described in this book. Finally, one can combine inadequate survey data on migration with data on income tax returns made available by the Internal Revenue Service (Engels & Healy, 1981). The survey data may yield adequate data on the age structures of total out- and in-migration counts but offer unreliable descriptions of spatial patterns of place-to-place flows. The IRS data may provide more accurate spatial structures of the migration flows. We explore such a situation in Chapter 6.

1.4 Outline of Book

The remaining six chapters of this book fall into three categories: models, applications, and a conclusion. Chapters 2 and 3 describe *models* that can be fitted to describe and summarize the structures of age patterns and of spatial patterns, respectively. Chapters 4, 5, and 6, respectively, focus on methods that apply these models to *smooth* irregular data, to *impose* structures on inadequate data, and to *infer* patterns of missing data. Finally, Chapter 7 offers concluding remarks and points to directions for further study.

Chapter 2
Describing Age Structures of Migration

2.1 Introduction

Empirical schedules of age-specific rates exhibit remarkably persistent regularities in age pattern. Mortality schedules, for example, normally show a moderately high death rate immediately after birth, after which the rates drop to a minimum between ages 10 and 15, then increase slowly until about age 50, and thereafter rise at an increasing pace until the last years of life. Fertility rates generally start to take on nonzero values at about age 15 and attain a maximum somewhere between ages 20 and 30; the curve is unimodal and declines to zero once again at some age close to 50. Similar unimodal profiles may be found in schedules of first marriage, divorce, and remarriage (Rogers, 1986). The most prominent regularity in age-specific schedules of migration is the high concentration of migration among young adults; rates of migration also are high among children, starting with a peak during the first year of life, dropping to a low point during the teenage years, turning sharply upward to a peak near ages 20–22, and then declining regularly thereafter, except for a possible slight hump at the onset of the principal ages of retirement, and/or an upward slope at the oldest ages.

We begin this chapter with an examination of regularities in age profile exhibited by empirical schedules of migration rates and go on to adopt the notion of model migration schedules to express these regularities in mathematical form. We then use model schedules to examine patterns of variation present in a large data bank of such schedules. Drawing on this comparative analysis of observed model schedules, we develop typologies and several "families" of schedules, and discuss the sensitivity of the model migration schedule to changes in one or more underlying parameters. We conclude by identifying a number of practical uses of model migration schedules.

2.2 Age Patterns of Migration

The simplest and most common measure of migration is the crude migration rate, defined as the ratio of the *number of migrants*, leaving a particular population located in space and time, to the average *number of persons* (more exactly,

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series
on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_2,

the number of person-years) exposed to the risk of becoming migrants. Data on non-surviving migrants are often unavailable, therefore the numerator in this ratio generally excludes them.

Because migration is highly age selective, with a large fraction of migrants being young, our understanding of migration patterns and dynamics is aided by computing migration rates for each single year of age. Summing these rates over all ages of life gives the *gross migraproduction rate (GMR),* the migration analog of fertility's gross reproduction rate. This rate reflects the level at which migration occurs out of a given region, and may be directional (i.e., from region *i* to region *j*).

2.2.1 Migration Rates and Migration Schedules

Age-specific migration schedules of multiregional populations exhibit remarkably persistent regularities. For example, when comparing the age-specific annual rates of residential migration among whites and blacks in the United States during 1966–1971, one finds a common profile (Fig. 2.1). Migration rates among infants and young children mirrored the relatively high rates of their parents, young adults in their late twenties. The mobility of adolescents was lower but exceeded that of young teens, with the latter showing a local low point around age 15. Thereafter migration rates increased, attaining a high peak at about age 22, and then declining monotonically with age to the ages of retirement. The migration *levels* of both whites and blacks were roughly similar, with whites showing a *GMR* of about 14 migrations and blacks one of approximately 15 migrations, over a lifetime undisturbed by mortality.

Although it has frequently been asserted that migration is strongly sex selective, with males being more mobile than females, recent research has indicated that sex selectivity is much less pronounced than age selectivity and is less uniform across time and space. Nevertheless, because most models and studies of population dynamics distinguish between the sexes, many migration measures do also.

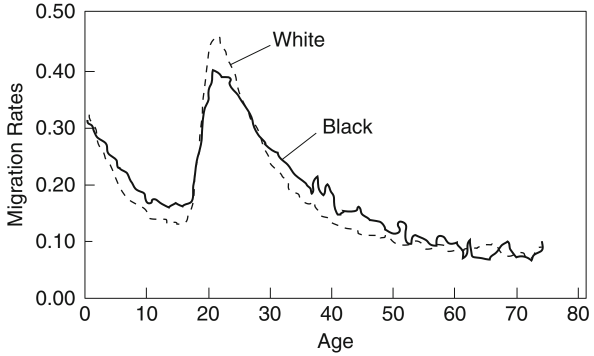


Fig. 2.1 Observed annual migration rates by race and single years of age: U.S., 1966–1971. (*Source*: Rogers & Castro, 1981)

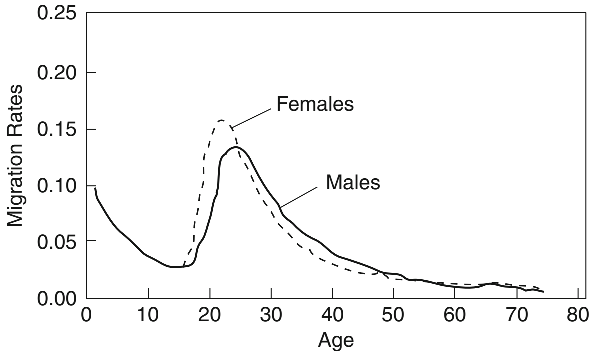


Fig. 2.2 Observed annual intercommunal migration rates by sex and single years of age: Sweden (average of annual rates, 1968–1973).
(*Source*: Rogers and Castro, 1981)

Figure 2.2 illustrates the age profiles of male and female migration schedules in Sweden. The migration levels are similar and the levels for males and females are roughly the same. The age profiles, however, show a distinct and consistent difference. The high peak of the female schedule precedes that of the male schedule by an amount that appears to approximate the difference between the average ages at marriage of the two sexes.

Under normal statistical conditions, point-to-point movements are aggregated into streams between one civil division and another; consequently, the level of interregional migration depends on the size of the areal unit selected. Thus if the areal unit chosen is a minor civil division such as a county or commune, a greater proportion of residential location will be included as migration than if the areal unit chosen is a major civil division such as a state or province.

Figure 2.3 presents the age profiles of U.S. female migration schedules as measured across different sizes of areal units: (1) all migrations from one residence to another, (2) changes of residence within county boundaries, (3) migration

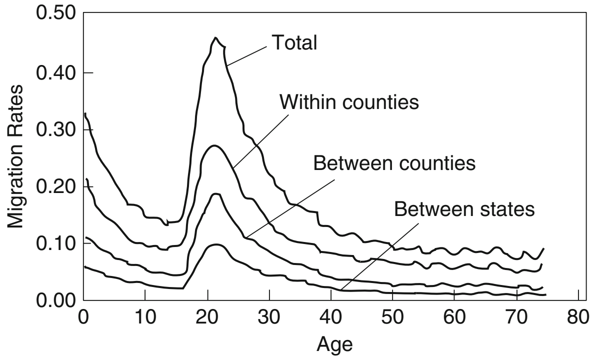


Fig. 2.3 Observed average annual migration rates of females by levels of area aggregation and single years of age: U.S., 1966–1971.
(*Source*: Rogers & Castro, 1981)

between counties, and (4) migration between states. The respective four *GMRs* are 14.3, 9.3, 5.0, and 2.5. The four age profiles appear to be remarkably similar, indicating that the regularity in age pattern persists across areal delineations of different size.

Finally, migration occurs over time as well as across space; therefore, studies of its patterns must trace its occurrence with respect to a time interval, as well as over a system of geographical areas. In general, the longer the time interval, the larger the number of return movers and nonsurviving migrants and, hence, the more the count of migrants will understate the number of interarea movers (and, of course, also of moves). Philip Rees, for example, after examining the ratios of 1- to 5-year migrants between the Standard Regions of Great Britain, found that

> . . .the number of migrants recorded over five years in an interregional flow varies from four times to two times the number of migrants recorded over one year. (Rees, 1977, p. 247)

This is the so-called *1-year/5-year problem* that we shall encounter when we study the migration patterns revealed by the American Community Survey, which asks a 1-year interval migration question, unlike the past five U.S. censuses which asked a 5-year interval question (1960–2000).

2.2.2 The Model Migration Schedule

The multiexponential (Rogers-Castro) model migration schedules described in this chapter are reduced forms of the 13-parameter expression that is comprised of five components:

1. A single negative exponential curve of the *pre-labor force ages*, with its descent parameter α_1;
2. A left-skewed unimodal curve of the *labor force* ages positioned around μ_2 on the age axis and exhibiting parameters of ascent λ_2 and descent α_2;
3. An almost bell-shaped curve of the *post-labor force* ages positioned around μ_3 on the age axis and exhibiting parameters of ascent λ_3 and descent α_3;
4. A single *positive* exponential curve of the post-retirement ages, with its ascent parameter λ_4;
5. A constant term, c.

The decomposition described above leads to the definition of the migration rates as the following simple sum of five components (Rogers and Castro, 1981, 1986; Rogers and Watkins, 1987; Rogers, 1988):

$$\begin{aligned} M(x) = {} & a_1 \exp(-\alpha_1 x) \\ & + a_2 \exp\{-\alpha_2 (x-\mu_2) - \exp[-\lambda_2 (x-\mu_2)]\} \\ & + a_3 \exp\{-\alpha_3 (x-\mu_3) - \exp[-\lambda_3 (x-\mu_3)]\} \\ & + a_4 \exp(\lambda_4 x) \\ & + c \end{aligned} \tag{2.1}$$

The *full* model schedule in Eq. (2.1) has 13 parameters. The profile of the full model schedule is defined by 8 of the 13 parameters: α_1, α_2, μ_2, λ_2, α_3, μ_3, λ_3 and λ_4; its level is determined by the remaining 5 parameters: a_1, a_2, a_3, a_4 and c. A change in the value of the area under a particular model schedule alters proportionally the values of the latter but does not affect the former.

In a comparative analysis of more than 500 migration schedules, Rogers and Castro (1981, 1986) identified a wide variety of age profiles, the most common of which was the 7-parameter reduced form of the model schedule, which consists of the first two components and the constant term. A significant number of schedules exhibited a pattern of migration in the post-labor force ages that followed the 11-parameter model migration schedule, in which the third "retirement" peak component was also present. In other schedules, instead of a retirement peak, the age profile took on the form of an "upward slope," where the fourth component replaced the third. In such instances, a 9-parameter basic model migration schedule was adopted. Finally, in a study of elderly migration, Rogers and Watkins (1987) found a number of instances when *both* a retirement peak and a post-retirement upward slope were exhibited, necessitating the full model schedule description with 13 parameters. In this chapter (and book), however, we only consider reduced forms of Eq. (2.1): the 7-, 9-, and 11-parameter versions.

The labor force and the post-labor force retirement components in Eq. (2.1) adopt the "double exponential" curve used by Coale and McNeil (1972) for their studies of nuptiality and fertility. The method chosen for fitting the model schedule to the data is a functional-minimization procedure known as the modified Levenberg-Marquardt algorithm (see Brown & Dennis, 1972, Levenberg, 1944, Marquardt, 1963). Minimum chi-square estimators were used in order to give more weight to age groups with smaller rates of migration. Table 2.1 sets out the estimated values for the basic and derived measures of the model migration schedule for Stockholm males illustrated in Fig. 2.4.

Model migration schedules of the form specified in Eq. (2.1) may be classified into *families* according to the ranges of values taken on by their principal parameters. For example, we may order schedules according to their migration levels, as defined by the values of the level parameters in Eq. (2.1), or by their associated *GMRs*. Alternatively, we may distinguish schedules with a retirement peak from those without one, or we may refer to schedules with relatively low or high values for the rate of ascent of the labor force curve λ_2, or the mean age. In many applications, it is also meaningful to characterize migration schedules in terms of several of the fundamental derived measures illustrated in Fig. 2.4, such as the low point x_l, the high peak x_h, and the retirement peak x_r. Associated with the first pair of points is the labor force shift X, which is defined to be the difference in years between the ages of the high peak and the low point, i.e., $X = x_h - x_l$. The increase in the migration rate of individuals aged x_h over those aged x_l is called the jump B.

The close correspondence between the migration rates of children and those of their parents suggests another important shift in observed migration schedules. If,

Table 2.1 Parameters and variables defining observed model migration schedules: Out-migration of males and females from the Stockholm region, 1974: Observed data by single years of age

	Stockholm	
Parameters and variables	Male	Female
GMR	1.45	1.43
a_1	0.042	0.041
α_1	0.097	0.091
a_2	0.059	0.067
μ_2	20.80	19.32
α_2	0.077	0.094
λ_2	0.374	0.369
a_3	0.000	0.000
μ_3	76.55	85.01
α_3	0.776	0.369
λ_3	0.145	0.072
c	0.003	0.003
$\bar{n}$	31.02	29.54
%(0–14)	25.61	25.95
%(15–64)	64.49	65.10
%(65+)	9.90	8.94
δ_{1c}	13.56	13.06
δ_{12}	0.716	0.604
δ_{32}	0.003	0.003
β_{12}	1.26	0.977
σ_2	4.86	3.94
σ_3	0.187	0.196
x_l	16.39	14.81
x_h	24.68	22.70
x_r	64.80	61.47
X	8.29	7.89
A	27.87	25.49
B	0.029	0.030

Source: Rogers and Castro (1981)

for each point x on the post-high-peak part of the migration curve, we obtain by interpolation the age (where it exists), $x - A_x$ say, with the identical rate of migration on the pre-low-point part of the migration curve, then the average of the values of A_x, calculated incrementally for the number of years between zero and the low point x_1, will be defined as the observed parental shift A.

An observed (or a graduated) age-specific migration schedule may be described in a number of useful ways. For example, references may be made to the heights at particular ages, to locations of important peaks or troughs, to slopes along the schedule's age profile, to ratios between particular heights and locations. The various

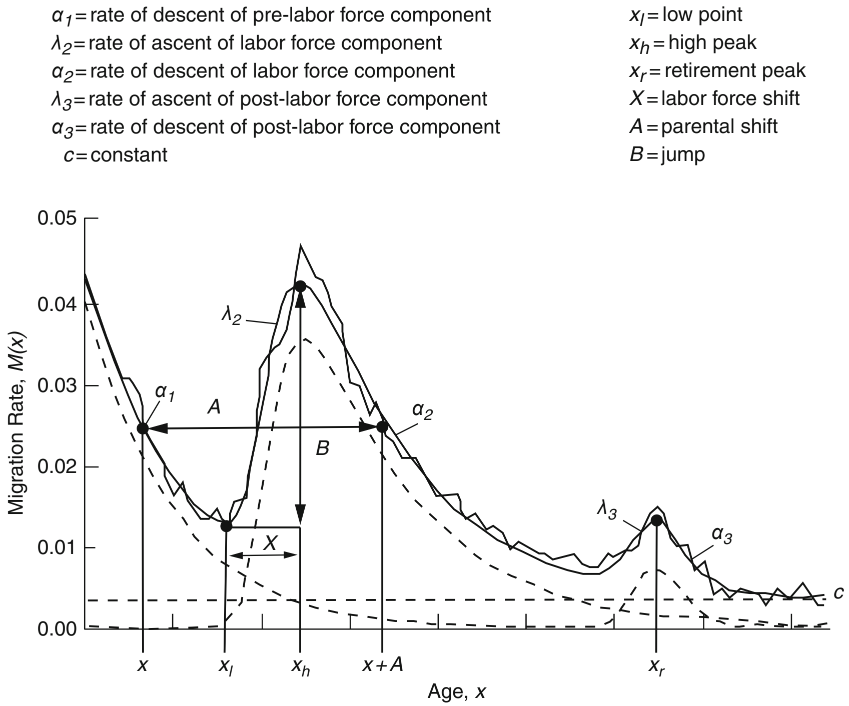


Fig. 2.4 The model migration schedule fitted to the observed out-migration rates of males leaving the Stockholm region, 1974.
(*Source*: Rogers & Castro, 1981)

descriptive measures characterizing an age-specific model migration schedule may be conveniently grouped into the following categories and subcategories:

1. *Basic measures* (in the 13 fundamental parameters and their ratios)

Heights:	a_1, a_2, a_3, a_4 and c
Locations:	μ_2, μ_3
Slopes:	$\alpha_1, \alpha_2, \lambda_2, \alpha_3, \lambda_3, \lambda_4$
Ratios:	$\delta_{1c} = a_1/c, \delta_{12} = a_1/a_2, \delta_{32} = a_3/a_2, \beta_{12} = \alpha_1/\alpha_2, \sigma_2 = \lambda_2/\alpha_2,$ $\sigma_3 = \lambda_3/\alpha_3$

2. *Derived measures* (properties of the model schedule)

Areas:	*GMR*, %(0-14), %(15-64), %(65+)
Locations:	$\bar{n}, x_1, x_h, x_r$
Distances:	X, A, B

2.3 Comparative Analysis

We've seen that observed age-specific rates of migration exhibit a fundamental age profile, which can be expressed in mathematical form as a model migration schedule defined by a total of seven to eleven parameters (we do not focus on the 13-parameter version in this book). Below, we examine and summarize the ranges of values typically assumed by each of these parameters and their associated derived variables calculated as part of a multinational Comparative Migration and Settlement Study (CMSS), carried out in the late 1970s and early 1980s at the International Institute for Applied Systems Analysis (IIASA) in Austria: (Rogers, 1978b; Rogers & Willekens, 1986; Willekens & Rogers, 1978).

IIASA's study of migration and settlement began with two basic components: a set of computer programs for multiregional demographic analysis and a network of collaborating investigators from nations of the Institute's then 17 member organizations. The principal goal was a case study of each country to be carried out by a scholar from that country. Each study used a common methodology and followed a common outline of substantive topics. Much of the data analysis was carried out at IIASA using a standard package of computer programs and most of the scholars involved had to be trained in the methodology by those at IIASA familiar with the mathematical theory.

The Migration and Settlement Study was concluded in 1982, seven years after its initiation. An important product of the study was the set of 17 country reports authored by 27 scholars. Each report presented a national overview of recent regional patterns of fertility, mortality, and internal migration, illustrated the application of multiregional demographic techniques and the insights into population redistribution that they revealed. An important by-product of the comparative study was the collection of interregional migration data that it acquired. That data set was analyzed in Rogers and Castro (1981, 1986), and some of that analysis is reported below.

2.3.1 An Example: The Swedish Case Study

The age-specific migration rates that were used to demonstrate the fits of the model migration schedule in the last section (Fig. 2.4) were single-year rates for a single-year time interval. Such data are scarce at the regional level and, in the IIASA comparative study, were available only for the eight-region disaggregation of Sweden (Andersson & Holmberg, 1980). But a comparison of the various parameter estimates for female schedules with those obtained when the same data were first aggregated to 5-year age groups and then disaggregated to single years of age by a cubic-spline interpolation indicated that such an interpolation procedure gives generally satisfactory results (see Table 2 in Rogers & Castro, 1981, pp. 14–15).

A number of useful measures describing the fitted model migration schedules for the out-migration rates from Stockholm to the Rest of Sweden are presented in Table 2.1. Some reflect levels of migrations, others pure age profile indicators. The former vary with levels of the *GMR*, the latter do not (hence the designation "pure"). Four parameters refer only to migration levels: a_1, a_2, a_3 and c.

The remaining model schedule parameters refer to the migration age profile: α_1, μ_2, α_2, λ_2, μ_3, α_3 and λ_3. Their values remain constant for all levels of the *GMR*. Taken together, they define the "pure" age profile of migration from one region to another. Schedules without a retirement peak yield only the four profile parameters:

α_1, μ_2, α_2 and λ_2, and schedules with a post-retirement slope add a single additional profile parameter λ_4.

Consider, for example, the ages of the low and high points in the pre-labor and young adult labor parts of the schedules. Males leaving Stockholm exhibit a low point of 16.39 years and a high point of 24.68 years. Females, on the other hand, show a low point that is younger (14.81 years) and a high point that also is younger (22.70 years). Retirement peaks center on age 64.80 for males and 61.47 for females, probably a reflection of the difference in the average ages at marriage. The average parental shift for males is 27.87 years and for females it is 25.49 years.

The contrasts identified above are rather typical for other fitted model schedules describing annual interregional migration age profiles in Sweden in 1974. The eight regions were defined in the Andersson and Holmberg (1980) report of IIASA's Migration and Settlement Study.

Rates of migration from each of eight regions to the rest of Sweden, if disaggregated by region of destination, gives $8^2 = 64$ directional schedules that need to be examined for each sex, which complicates comparisons with other nations. To resolve this difficulty, Rogers and Castro (1981, 1986) associated a "typical" schedule with each collection of national rates by calculating the mean of each parameter and derived variable.

To avoid the influence of unrepresentative "outlier" observations in the computations of averages defining a typical national schedule, it was decided to delete approximately 10% of the "extreme" schedules. Specifically, the parameters and derived variables were ordered from low value to high value; the lowest 5% and the highest 5% were defined to be extreme values. Schedules with the largest number of low and high extreme values were discarded, in sequence, until only about 90% of the original number of schedules remained. This *reduced* set then served as the population of schedules for the calculation of various summary statistics. Table 2.2 illustrates the average pure profile parameter values obtained with the Swedish data. (Since the median, mode, standard deviation-to-mean ratio, and lower and upper bounds were also of interest, they were included as part of the more detailed computer outputs reproduced in Appendix B of Rogers and Castro (1981).

Table 2.2 presents information about patterns of migration by age. The parameters, given in columns, define a range of fitted model migration schedules. In general, their particular pure profile parameter estimates and derived variable values are similar to those associated with the out-migration schedules for Stockholm (Table 2.1). The low point averages range from 14.44 to 16.49 years and the corresponding high point values from 21.72 to 24.46 years. Once again the high point values for females are lower than for males, as are the values of the parental shift.

2.3.2 Families of Schedules: Toward a Typology

One can imagine describing a model migration schedule along its vertical and horizontal dimension. For example, the heights of the labor force and pre-labor force components are reflected in the parameters a_2 and a_1, respectively, therefore the ratio a_2/a_1 indicates the degree of the "labor dominance," and its reciprocal,

Table 2.2 Mean values of pure profile parameters and derived variables defining the reduced set of observed model migration schedules: Sweden, 8 regions, 1974 observed data by single years of age[a]

	Males		Females	
Parameters	Without retirement peak (48 schedules)	With retirement peak (9 schedules)	Without retirement peak (54 schedules)	With retirement peak (3 schedules)
α_1	0.124	0.085	0.108	0.093
μ_2	20.502	21.249	19.094	18.868
α_2	0.104	0.093	0.127	0.106
λ_2	0.448	0.416	0.537	0.424
μ_3		76.711		74.781
α_3		0.847		0.938
λ_3		0.158		0.170
x_l	15.621	16.494	15.260	14.444
x_h	23.571	24.462	21.720	21.904
x_r		65.630		64.604
X	7.950	7.968	6.461	7.460
A	30.270	28.668	27.222	26.119
B	0.030	0.024	0.036	0.026

[a] Region 1 (Stockholm) is a single-commune region; hence there exists no intraregional schedule for it, leaving $8^2 - 1 = 63$ schedules, of which 6 were deleted
Source: Rogers and Castro (1981)

$\delta_{12} = a_1/a_2$, the index of "child dependency." The lower the value of δ_{12}, the lower the degree of child dependency exhibited by a migration schedule and, correspondingly, the greater its labor dominance. This suggests a dichotomous classification of migration schedules into *child dependent* and *labor dominant* categories.

Labor dominance reflects the relative migration levels of those in the working ages relative to those of children. Labor asymmetry, on the other hand, refers to the shape of the left-skewed unimodal curve describing the age profile of labor force migration. Imagine that a perpendicular line, connecting the high peak with the base of the bell-shaped curve (i.e., the jump B), divides the base into two segments g and h. Clearly, the ratio *h/g* is an indicator of the degree of asymmetry of the curve. A more convenient index, using only two parameters of the model schedule is the ratio $\sigma_2 = \lambda_2/\alpha_2$, the index of labor asymmetry. Its movement is highly correlated with that of *h/g,* because of the approximate relation

$$\sigma_2 = \lambda_2/\alpha_2 \propto \frac{B/g}{B/h} = h/g \tag{2.2}$$

where $\propto$ denotes proportionality. Thus σ_2 may be used to classify migration schedules according to their degree of labor asymmetry. Again, an analogous argument applies to the post-labor force curve, and $\sigma_3 = \lambda_3/\alpha_3$ may be defined as the index of retirement asymmetry.

When "adding on" a pre-labor force curve of a given *level* to the labor force component, it is also important to indicate something of its *shape*. For example, if the migration rates of children mirror those of their parents, then α_1 should be approximately equal to α_2 and $\beta_{12} = \alpha_1/\alpha_2$, the index of parental-shift regularity, should be close to unity.

Large differences in the *GMR*s of fitted model migration schedules, however, give rise to slopes and vertical relationships that are not comparable when examined visually. Recourse then must be made to a standardization of the areas under the migration curves, for example, a general rescaling to a *GMR* of unity. Recall that this difficulty does not arise when comparing values of the principal slope and location parameters and their ratios, because such indices when used to characterize the schedules are not affected by changes in levels. Only heights, areas, and vertical distance measures, such as the *jump*, are level-dependent measures.

The analysis of the collection of age-specific directional migration schedules described in Rogers and Castro (1981, 1986) suggests that a useful typology of such schedules can be developed by simply cross-classifying high and low levels of mobility with early and late peaking (timing) of mobility, the "how much" and the "when" of migration. Associated with these four types for out-migration are four important points on the prototypical age pattern of migration: the infant migration level, the age at which migration hits a low point during the teenage years and begins its ascent, the age at which the associated high point (called the labor force peak) occurs and, finally, the age profile of migration of the older post-labor force ages, the presence or absence of either a retirement peak or of a gradual rise at the oldest ages or both (Rogers and Castro, 1981; Rogers and Watkins, 1987).

The height of the infant migration peak has been shown to be associated with the level of fertility in the population (Castro & Rogers, 1983). The low point during the teenage years and the start of the ascent to the labor force peak is related to the population's average age of leaving home. The location and relative height of the labor force peak is associated with the pace at which teens and young adults leave home to enter the labor market, go to college, enroll in the military, and get married. And finally, the presence of a retirement peak typically only occurs in more developed countries in migration flows directed from cold to warm climates (e.g., from New York to Florida). Another variable of interest is the *parental-shift* denoted by *A* in Fig. 2.4, which Rogers and Castro (1981) define as the average number of years separating the migration rates of children from those of their parents with whom they must migrate, a value that normally is close to the average age of childbearing.

The comparative analysis of national and interregional migration patterns carried out in Rogers and Castro (1981) identified several distinct *families* of age profiles. Within each family of schedules, a number of key parameters and variables were put forward to further classify different categories of migration profiles. For example, the following aspects of shape and location along the age axis were found to be important:

(1) *Peaking*: early versus late peaking (μ_2)
(2) *Dominance*: child dependency (δ_{12}) versus labor dominance ($\delta_{21} = 1/\delta_{21}$)
(3) *Asymmetry*: labor symmetry versus labor asymmetry (σ_2)
(4) *Regularity*: the degree to which the migration rates of children mirror those of their parents (β_{12}).

Three sets of model migration schedules have been studied in this chapter: the 11-parameter schedule with a retirement peak, the alternative 9-parameter schedule with a retirement slope, and the simple 7-parameter schedule with neither a peak nor a slope. Thus we have at least three broad families of schedules. Not examined here is the full 13-parameter schedule; but all four families are illustrated in Fig. 2.5 below.

Additional dimensions for classifying schedules into families are suggested by the above comparative analysis of national migration age profiles and the basic measures and derived variables defined in Section 2.2. These dimensions reflect different locations on the horizontal and vertical axes of the schedule, as well as different ratios of slopes and heights.

We may imagine a 3 × 4 cross-classification of migration schedules that defines a dozen "average families" with measures that take on the average values found in the more than 500 migration schedules examined in Rogers and Castro (1981, 1986). Introducing, in addition, a low and high value for each of the four basic measures gives rise to additional families for each of the three classes of schedules. Thus we may conceive of a collection of families, divided into schedules with a retirement peak, schedules with a retirement slope, and schedules with both (or neither).

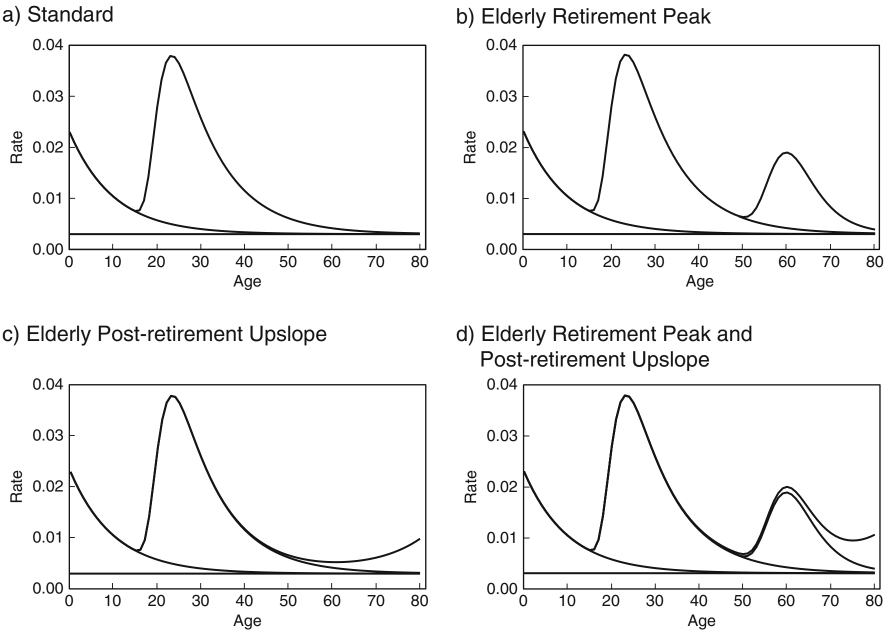


Fig. 2.5 The four main families of Rogers-Castro model migration schedules: standard, elderly retirement peak, elderly post-retirement upslope, and elderly retirement peak and post-retirement upslope.
(*Source*: Raymer & Rogers, 2008)

2.4 Related Topics

2.4.1 Sensitivity Analysis

The preceding section has focused on a comparison of the fundamental parameters defining the model migration age profiles of a number of nations. The comparison yielded ranges of values within which each parameter may be expected to fall and suggested a classification of schedules into families. We now turn to an analytic examination of how changes in several of the more important parameters become manifested in the age profile of the model schedule. For analytical convenience we begin by focusing on the properties of the double exponential curve that describes the labor force component:

$$f_2(x) = a_2 \exp\{-\alpha_2 (x - \mu_2) - \exp[-\lambda_2 (x - \mu_2)]\} \tag{2.3}$$

We begin by observing that if α_2 is set equal to λ_2 in the above expression then the labor force component assumes the shape of a well-known extreme value distribution used in the study of flood flows (Gumbel 1941; Kimball 1956). In such a case $x_h = \mu_2$ and the function $f_2(x)$ achieves its maximum y_h at that point. To analyze the more general case where $\alpha_2 \neq \lambda_2$, we may derive analytical expressions for both of these variables by differentiating Eq. (2.3) with respect to x, setting the result equal to zero, and then solving to find

$$x_h = \mu_2 - (1/\lambda_2) \ln(\alpha_2/\lambda_2) \tag{2.4}$$

an expression that does not involve a_2, and

$$y_h = a_2 (\alpha_2/\lambda_2)^{\alpha_2/\lambda_2} \exp(-\alpha_2/\lambda_2) \tag{2.5}$$

an expression that does not involve μ_2. Note that if $\lambda_2 > \alpha_2$, which is almost always the case, then $x_h > \mu_2$. And observe that if $\alpha_2 = \lambda_2$, then the above two equations simplify to $x_h = \mu_2$ and $y_h = a_2/e$.

Since μ_2 affects x_h only as a displacement, we may focus on the variation of x_h as a function of α_2 and λ_2. A plot of x_h against α_2, for a fixed λ_2, shows that increases in α_2 lead to decreases in x_h. Analogously, increases in λ_2, for a fixed α_2, produce increases in x_h but at a rate that decreases rapidly as the latter variable approaches its asymptote.

The behavior of y_h is independent of μ_2 and varies proportionately with a_2. Hence its variation also depends fundamentally only on the two variables α_2 and λ_2. A plot of y_h against α_2, for a fixed λ_2, gives rise to a U-shaped curve that reaches its minimum at $\alpha_2 = \lambda_2$. Increasing λ_2 widens the shape of the U.

The introduction of the pre-labor force component into the profile generally moves x_h to a slightly younger age and raises y_h by about $a_1 \exp(-\alpha_1 x_h)$, usually a negligible quantity. The addition of the constant term c, of course, affects only y_h, raising it by the amount of the constant. Thus the migration rate at age x_h may be expressed as

$$M(x_h) \approx a_1 \exp(-\alpha_1 x_h) + y_h + c \tag{2.6}$$

A variable that interrelates the pre-labor force and labor force components is the parental shift A. To simplify our analysis of its dependence on the fundamental parameters, it is convenient to assume that α_1 and α_2 are approximately equal. In such instances, for ages immediately following the high peak x_h, the labor force component of the model migration schedule is closely approximated by the function $a_2 \exp[-\alpha_2 (x - \mu_2)]$. Recalling that the pre-labor force curve is given by $a_1 \exp(-\alpha_2 x_1) + y_h + c$ when $\alpha_1 = \alpha_2$, we may equate the two functions to solve for the difference in ages that we have called the parental shift:

$$A = x_2 - x_1 = \mu_2 + (1/\alpha_2) \ln(1/\delta_{12}) \tag{2.7}$$

This equation shows that the parental shift will increase with increasing values of μ_2 and will decrease with increasing values of α_2 and δ_{12}. A comparison of the values of this analytically defined "theoretical" parental shift with the corresponding observed parental shifts obtained for males and females in eight Swedish regions produced similar numerical values; the analytical definition has the advantage of being simpler to calculate and analyze, but it is very sensitive and depends on "good" estimates of the model schedule parameters (see Table 14 in Rogers & Castro, 1981, p. 33).

2.4.2 The 1-Year/5-Year Problem

In this section, we examine the migration flows that are represented by the conditional survivorship proportions of persons migrating from origin i to destination j for 1-year migration time intervals (i.e., current place of residence by place of residence one year ago) and 5-year migration time intervals (i.e., current place of residence by place of residence five years ago). Recall that conditional survivorship proportions represent the surviving proportions of persons who migrated from origin i to destination j. For each origin region, they define the proportion staying and the proportion migrating. Such proportions are said to be conditional because only those who survived to the time of the census could report their migration status.

A number of analysts have looked into the problem of reconciling 1- and 5-year migration data, (Rees, 1977; Kitsul & Philipov, 1982). Rogerson (1990) argues not only that the level of mobility but also the geographic pattern of migration flows is influenced by the choice of interval width. He notes that the return and onward migrations that occur within the 5-year period, along with the heterogeneity in the flows are responsible for creating the differences.

Rogers, Raymer, and Newbold (2003) review and re-examine differences in *level* and *spatial patterns* finding 5- to 1-year ratios of levels that vary between 3 and 3.5 and spatial allocations of those levels that exhibit a surprising degree of relative stability. Here we introduce a consideration of changing *age patterns*. Specifically, how to the age profiles of migration identified by 1-year and 5-year time intervals differ?

Figure 2.6 sets out the model schedule age profiles of two *sets* of U.S. migration data. The differences between the 1-year ACS data and the 5-year Census data are evident.

A motivation for our examination of the problem comes partly from the changes currently occurring with regard to migration data collected in the U.S. In the future, data on internal migration collected by the U.S. Census Bureau will only come from the American Community Survey (ACS), which is a continuing monthly survey that will replace the historical census long-form by the year 2010. The data from the ACS have the advantages of lower cost and more up-to-date information, but this replacement of the long-form questionnaire has lead to new perspectives on the measurement and analysis of internal migration flows, and on the development of intercensal population estimates.

The new migration data reflects place of residence one year ago, though not at one fixed point in time, but rather the result of continuous sampling throughout the

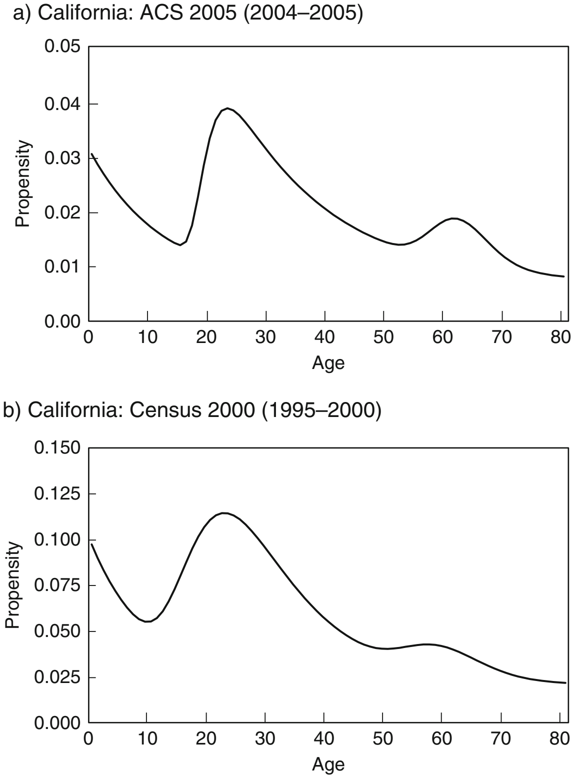


Fig. 2.6 The model schedule age profiles of California's out-migration conditional survivorship proportions: 1-year time interval (ACS 2005) versus 5-year time interval (Census 2000)

course of each year. In contrast, in the past five censuses, the migration data has reflected place of residence exactly five years ago. This change creates problems for those users of the ACS data who need information on migration to calibrate cohort-survival population projection models with age and time intervals of 5-years. It also directs increased attention at the underlying differences between counts of migrants versus counts of migrations.

Migration flow data collected over time intervals of differing lengths differ in their measures of migrations and migrants. By missing some return and onward migrations, the longer intervals undercount the actual number of moves, and thereby tend to emphasize the movement levels and spatial patterns of those who move less frequently. Moreover, every migration flow can be disaggregated into its primary, return, and onward components. Typically, each of these three subgroups exhibits different propensities to migrate, different spatial patterns, and different age profiles. Since the relative representation of each of these three subgroups in the observed total flow is influenced by the width of the time interval selected (and the spatial scale of the regionalization), clearly the associated migration levels, spatial structures, and age patterns will differ.

2.4.3 The Age Composition of Migrants

The age profile of a population is often used as evidence of the historical fertility and mortality patterns that gave rise to it. For example, Fig. 2.7, Panel (a), shows the contrast in the age composition of the populations of Mexico in 1970 and of Sweden in 1974 (Castro & Rogers, 1983). The Mexican population distribution suggests high rates of natural increase and mortality, increasing fairly rapidly with age. The Swedish data, on the other hand, illustrate a population distribution 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, Panel (b) of Fig. 2.7 reveals the age composition of internal migrants for Mexico and Sweden, and it is clear that the two population structures affect the migration age profiles. In Mexico the youngest age groups are a very large portion of the population, which is consistent with the high fertility rate, and it follows logically that the youngest age groups comprise a substantial portion of the migrating population. Sweden, on the other hand, exhibits a more uniform population distribution (Panel (a) of Fig. 2.7), and its age profile of migrants (Panel (b)) reflects the relative propensities to migrate that are more common among developed societies.

The data for Mexico and Sweden offer a visual illustration of a relationship between a population and its distribution of out-migrants, but that relationship may be set out analytically. Age-specific out-migration rates are often denoted by $M(x)$, which is defined as $M(x) = O(x)/K(x)$, where $O(x)$ is the total number of out-migrants age x, and $K(x)$ is the total population at the same age. Rogers (1978b) showed that the age profile of $M(x)$ is related to the distribution of out-migrants and the distribution of the population from which they were a part. This can be demonstrated using the definition of $M(x)$. If we denote the proportion of out-migrants

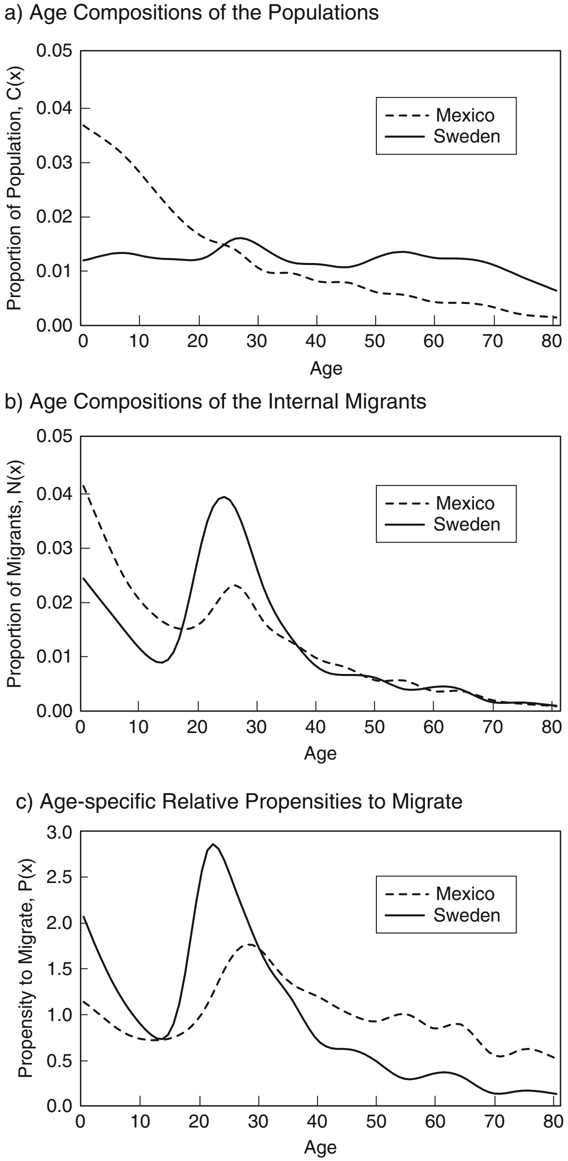


Fig. 2.7 Two sets of national age composition profiles: Mexico, 1970, and Sweden, 1974. (*Source*: Little & Rogers, 2007, Fig. 2.1, adapted from Castro & Rogers, 1983)

aged x as $N(x)$, and the proportion of the total population aged x as $C(x)$, then the definition of $M(x)$ can be expressed as a function of $N(x)$ and $C(x)$:

$$M(x) = \frac{O(x)}{K(x)} = \frac{O^*N(x)}{K^*C(x)}, \tag{2.8}$$

where O is the total number of out-migrants and K is the total population number (Castro & Rogers, 1983). Then Eq. (2.10) can be rearranged so that the age composition of out-migrants $N(x)$ is expressed as a function of two distributions:

$$N(x) = C(x)\left(\frac{M(x)}{O/K}\right) \tag{2.9}$$

and if

$$P(x) = \frac{M(x)}{O/K}$$

then

$$N(x) = C(x)^{*}P(x). \tag{2.10}$$

2.5 Summary and Discussion

This chapter began with the observation that empirical regularities characterize observed migration schedules in ways that are no less important than the corresponding well-established regularities in observed fertility or mortality schedules. The data analyzed confirmed that, although migration levels vary substantially from place to place, the shape of an age-specific schedule of migration propensities seems to be quite similar across a wide range of regions. Young adults in their early twenties generally exhibit the highest regional outmigration rates and young teenagers show the lowest. Because children migrate with their parents, infant migration rates are higher than those of adolescents. And retirement migration may give rise to a bell-shaped protrusion in the migration age profile around the ages of retirement.

Section 2.2 was devoted to defining mathematically regularities in observed migration schedules in order to exploit the notational, computational, and analytical advantages that such a formulation provides. Section 2.3 reported on the results of an examination of the migration schedules for a large number of countries. Section 2.4 focused on three related topics: the ways in which model migration schedules are shaped by the parameters of the model schedules, the problems of reconciling 1-year time interval data with 5-year interval data, and the representation of the age composition profiles of migrants. Later chapters will show how regularities in migration age profiles lead naturally to the development of model migration schedules that might be suitable for studies of populations with irregular, inadequate, or missing data.

Of what use, then, is the model migration schedule defined in this chapter? What are some of its concrete practical applications? The model migration schedule may be used to *graduate* observed data, thereby smoothing out irregularities and ascribing to the data summary measures that can be used for comparative analysis. It may be used to *interpolate* to single years of age, observed migration schedules that are reported for wider age intervals. Assessments of the *reliability* of empirical migration data and indication of appropriate strategies for their correction are aided by the availability of standard families of migration schedules that can be *imposed* on *unreliable data*. And such schedules also may be used to help resolve problems caused by *missing data*.

Finally, the application (i.e., fitting) of the model migration schedules to observed data requires the estimation of each schedule's parameters. These parameters have been estimated using standard non-linear estimation procedures. The Appendix to this chapter identifies such procedures and the computer software that was used to implement them.

2.6 Appendix: Estimation of Model Schedule Parameters

This chapter has focused on fitting observed data on age-specific migration propensities using the Rogers-Castro multiexponential model migration schedules. All of the estimates were obtained using one of three *non-linear* curve-fitting computer programs: MODEL, MODELMATLAB, and TableCurve 2D. The first was written in FORTRAN by research colleagues of Andrei Rogers: Luis Castro and Friedrich Planck at the International Institute of Applied Systems Analysis during the late 1970s and early 1980s, with revisions introduced by Jani Little at the University of Colorado, Boulder in the late 1990s. A brief description of MODEL may be found in Rogers and Little (1994). The second was written in MATLAB by Avleen Bijral and Jani Little, at the University of Colorado, Boulder in 2005–2006 and is used in Chapters 4, 5, and 6, when processing large numbers of required fittings. The third program used was a commercially offered general curve-fitting program purchased from Jandel Scientific, TableCurve 2D. (For an application, see Rogers & Raymer, 1999).

TableCurve 2D (version 5.0) is an automated curve fitting and equation discovery program that has been designed for a variety of scientific uses. In this program, the Rogers-Castro model migration schedule can be specified using a User Defined Function that allows up to 10 parameters to be estimated. This program also has visualization features that permit the user to see how the model is affected by changes in individual parameter values. Rogers and Raymer (1999) compared the results and procedures of this software with those of MODEL and found that both programs produced the same results, but that TableCurve 2D had several advantages, particularly the procedural aspects of the modeling process, which are more user-friendly. For example, imagine that a problem arises in the specified initial estimates, which is quite common when dealing with so many parameters. In TableCurve 2D, it is possible to partition the data and then to derive initial estimates for different sections of the curve. Levin and Mitra (1994) demonstrated this with the TableCurve 2D program using mortality data. Note, most standard statistical software (e.g., SPSS or Stata) have non-linear regression routines. These also can be used to estimate the 7-parameter schedule, however one needs reasonable initial estimates. Fitting the 9-, 11-, or 13-parameter schedules is much more complicated. Here, a graphical interface like the one in TableCurve 2D is very useful for obtaining the initial estimates.

Common to all three computer programs is the classical problem of *non-linear* parameter estimation in unconstrained optimization. All start with a set of initial guesses of the desired parameter values and then seek to improve the goodness-of-fit by identifying "better" values, until specific convergence criteria are met. This iterative sequence ends after a finite number of iterations, and the last set of estimates is accepted as giving the best fit of the multiexponential function to the observed data.

Except for the data in the Swedish case study, described in Section 2.3.1 and the American Community Survey (ACS) data described in later chapters, all age-specific migration rates or propensities in this book begin with the migration of 5-year age groups. We have found that more accurate parameter estimates are obtained if these latter data are first graduated to produce 1-year age group data using a cubic-spline interpolation procedure (McNeil, Trussell, & Turner, 1977).

Finally, the principal difficulty in non-linear parameter estimation is that of convergence. The algorithm begins by assuming a set of initial parameter values and ultimately ends with a set of "optimum" values. But the optimum may be merely a local optimum, and not the global *optimum optimorum*. A better guess of the initial parameter values may produce an improved goodness-of-fit and produce a different set of final values.

How to choose a "good" set of initial values? An effective procedure is to carry out a linear estimation method first, which does not rely on an iterative algorithm. That method was first described in Castro and Rogers (1981), applied and analyzed by Watkins (1984), and ultimately published as one of the several alternatives set out in Rogers et al. (2005).

Chapter 3
Describing Spatial Structures of Migration

3.1 Introduction

The notion of age structure is a central concept in demography, but the structure of migration, which is inherently spatial, is not commonly presented. The former has been used to develop functional representations of the age patterns of a population or that of a stream of migrants, and it is the basis for the construction of model migration schedules, mathematical expressions such as those that describe the age patterns of migration propensities in Chapter 2. The latter, on the other hand, has no such widely accepted mathematical representation. Yet it clearly exists, as the spatial pattern of the principal U.S. elderly retirement flows depicted in Fig. 3.1 illustrates. We offer such a definition, one that draws on Rogers et al. (2002) and the log-linear specification of the geographer's spatial interaction model.

As a demographic process, migration stands apart from fertility and mortality because of the explicitly spatial nature of migration. Unlike fertility and mortality processes, which affect the population of only one region, aggregate migration flows interact within a multiregional system in which departures from each region affect the populations of several other regions, subtracting people from each region of origin and adding people differentially to each region of destination. Therefore, representation of this complex process and associated data structure must come from a model that incorporates the influences of population sizes at the origins and destinations, and one that also includes some sort of "separation" or "interaction" factor between each pair of origins and destinations.

We define *migration spatial structure* to be a particular description of a matrix of interregional migration flows, one that provides an analyst with the means to: (1) reconstruct that matrix of flows from a set of parameters, (2) identify the implied relative "push" at each origin and "pull" of each destination, and (3) express the origin-destination-specific levels of spatial interaction implied by that matrix of flows. Spatial interaction is here taken to reflect the degree of deviation exhibited by the flow matrix when compared to the corresponding matrix generated under assumption of no spatial interaction, i.e., a situation in which origin-destination-specific migration flows, rates, or probabilities are independent of origin and destination; the larger the deviation the stronger the degree of spatial interaction.

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_3,

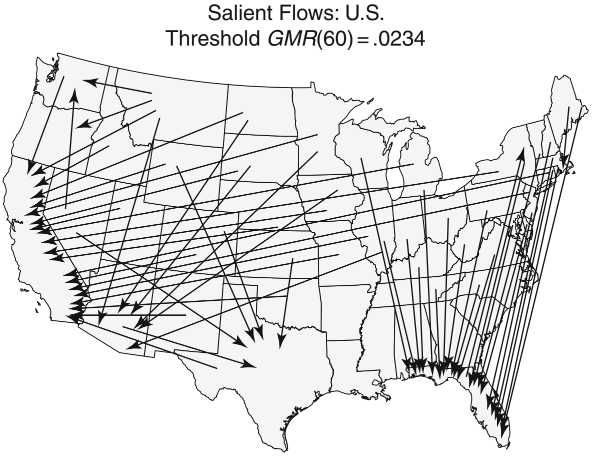


Fig. 3.1 The principal elderly retirement migration flows in the U.S., 1975–1980, excluding Alaska, Hawaii, and Washington, DC. (*Source*: Rogers et al., 1990, p. 270)

The linkage between age structure and the analysis of fertility, mortality, and migration processes is central to demographic study. Standardized mathematical representations of age patterns have allowed demographers accurately to define age-specific patterns with continuous functions described with relatively few parameters. The corresponding mathematical representation of spatial patterns calls for a somewhat more complex statistical structure. A powerful, yet conceptually simple, instrument for the study of aggregate migration spatial structure is offered by the family of generalized linear models, particularly the *log-linear model*. It provides a mathematical representation of migration flow structure that is more readily interpretable than is the flow matrix itself. Just as model schedules are used to make comparisons across time and place, the log-linear specification can be employed as a statistical model that is especially valuable for comparing interregional migration structures across time. The parameters of the log-linear model can be used to not only gauge the relative "push" and "pull" of specific regions but also the level of interaction (or association) between pairs of regions. Because the parameters of the model are interpretable and can be used to characterize migration spatial structure, the log-linear model has the potential for standardizing and enhancing demographic analysis. As we show in this book, the log-linear model, like model migration schedules, can be used to *smooth*, *impose* or *infer* migration flows.

The main contribution of the log-linear approach is that the parameters of that model capture different features of the spatial structure of migration, with one set of parameters representing the effect of the sizes of origin populations, another set representing the corresponding effects of the sizes of destination populations, and still another set representing the strengths of the linkages between these two populations. This parameterization facilitates comparisons of spatial structures. The method can be applied to a multiregional system comprised of many regions and to theoretical as well as observed spatial structures. It also decomposes the spatial structure into contributing structural factors. For example, the number of migrants from a region, *i* say, to another region, *j* say, depends on the size and composition of the population of region *i* and on the size and composition of the population of region *j*.

Population sizes and compositions alone, however, are not sufficient to characterize the flows of migrants. The spatial interaction between the populations of regions *i* and *j* is also important and, indeed, a history of migration from *i* to *j* may be a more important determinant of current migration structure than the particular characteristics of the two regions. In Section 3.3, a method is presented that is able to capture the effect on contemporary migration of historical migration patterns, facilitating the quantitative assessment of historical changes among observed structures and their influences on contemporary spatial structures. That method is the *method of offsets* (Knudsen, 1992). A particularly interesting observation is that this method belongs to the family of log-linear models.

We begin below with a brief overview of previous efforts to describe the spatial structure of migration. We then focus on the merits of the general spatial interaction model and emphasize the functional equivalence between this model and the log-linear model. We also offer an exposition of the log-linear model, showing how it can be used to represent the components of migration spatial structure, illustrating its use with particular numerical examples, and then considering its extensions and wider implications. Finally, we end the chapter with a summary and discussion.

3.2 Representing Spatial Structures of Migration: The Log-Linear Model

3.2.1 Overview

The literature on migration is curiously ambiguous on the subject of what is migration spatial structure and how it should be measured. An early effort to describe the structure of migration was that of Shryock (1964, p. 267) who put forward a preference index that focused on the ratio of actual to expected number of migrants in a stream, the latter defined as being proportionate to both the population at origin and the population at destination. Clayton (1977, p. 109), on the other hand, defined migration spatial structure as the way in which origins and destinations are linked in terms of their exchanges of migrants. He then implemented this definition by identifying those regions (states in his application) that acted as major origins and destinations in the interstate migration system. He used nodal and principal component analyses to identify such places and delineated a number of migration fields.

Plane (1984) and Manson and Groop (1996), among many others, relied on the widely used notion of migration *efficiency*, a measure of redistributional effectiveness, and applied it to interstate migration matrices to identify changes in migration system structure. And in a co-authored article with Mulligan, Plane adopted the well-known Gini index of concentration to identify the *spatial focus* exhibited by a set of origin-destination-specific migration flows, measuring the strength of the concentration by the departure from equality in the distribution of migration streams that is exhibited by an observed origin-destination-specific matrix of flows (Plane & Mulligan, 1997). Rogers and Sweeney (1998) and Rogers and Raymer (1998) instead focused on the coefficient of variation as the relevant index.

Finally, Mueser (1989) fitted a generalized spatial interaction model to data on migration flows between U.S. states over three decades. Mueser's work is important because he demonstrates the ability of the spatial interaction model to clearly represent the structural components of migration. Due to his reliance on the spatial interaction model, he was able to decompose migration structure into the "sending" effects of each region, the region's ability to "draw" migrants, and the inter-regional interaction or separation effects. His findings on migration flow stability conflict with the conclusions of Plane (1984) and Manson and Groop (1996), however, probably because of the differences in methodological approaches. Instead of instability, he finds that there is great stability in the separation effects, i.e., the relative attachments between regions over time. There are changes, he concludes, in the relative volumes of migration streams, but these are due to the relative desirability or draw of different locations rather than to the spatial interactions between them.

3.2.2 The Spatial Interaction Model and the Log-Linear Model

The spatial interaction model, once so popular in human geography (Haynes & Fotheringham, 1984), has proven to be the most useful method for representing the spatial structure of migration (Willekens, 1983a; Mueser, 1989). Its generality incorporates most models used to examine migration streams including the gravity model, entropy maximization, information minimization, biproportional adjustment, the systemic model of movement, random utility models based on choice theory, and the log-linear model. A formal equivalence exists between the log-linear model and the gravity model and entropy maximization model (e.g., see Willekens, 1980; 1982a; 1982b; 1983a; Bennett & Haining, 1985; Aufhauser & Fischer, 1985).

The log-linear model is a powerful instrument for the study of complex data structures. Its use to express traditional models of spatial interaction enhances the opportunities for structural analysis. Questions that the data are expected to help answer may be expressed in terms of the parameters of the model. Furthermore, the model clarifies and simplifies the *estimation* of spatial interaction flows. And when particular interaction effects cannot be derived from available data, they often may be calculated using other comparable data sets (e.g., historical data on interaction). Since Snickars and Weibull (1977) found that migration tables of the past provide much better estimates of current accessibility than any distance measure, historical data are often used in spatial interaction analysis to capture spatial patterns of accessibility. A drawback of using historical information is that this assumes that spatial interactions remain stable, i.e., that migration regimes are fixed. However, research on matrix transformation methods and log-rate models for representing past age and spatial patterns of structural change has provided us with a logical way to relax the strict assumption of an unchanging regime (Rogers & Taylor, 1996; Lin, 1999b).

3.2.3 Numerical Examples of the Log-Linear Decomposition

To illustrate the advantages of analyzing migration in terms of multiplicative components, consider the U.S. migration flows between the four Census Bureau-defined regions during the 1985–1990 and 1995–2000 time periods set out in Table 3.1. All persons who died, were born, or left or entered the country during the period have been excluded. Also, persons who remained in region i (i.e. non-migrants denoted n_{ii}) are excluded from the table. During the 1985–1990 period, 10.5 million persons over the age of 5 years were classified as interregional migrants. The Northeast and Midwest regions sent about half of all migrants but only received one third. The largest origin-destination-specific flow was from the Midwest to the South. How can we describe and compare the migration spatial structures exhibited by these flow matrices, using a spatial interaction model?

The elements (n_{ij}) in each of the two migration flow tables in Table 3.1 can be expressed as follows:

$$n_{ij} = (T)(O_i)(D_j)(OD_{ij}) \tag{3.1}$$

where n_{ij} is an observed flow of migration from region i to region j. This general type of model is called a *multiplicative component model*. Such a specification of the model is consistent with that of the log-linear model, that is, taking the natural logarithm of n_{ij} results in the corresponding additive model:

$$\ln(n_{ij}) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_{ij}^{OD} \tag{3.2}$$

Table 3.1 U.S. interregional migration flows (in thousands), 1985–1990 and 1995–2000

		Destination				
Period	Origin	Northeast	Midwest	South	West	Total
(a) 1985–1990	Northeast		330	1,604	469	2,403
	Midwest	344		1,672	947	2,963
	South	760	1,180		1,122	3,062
	West	340	658	1,053		2,051
	Total	1,444	2,168	4,329	2,538	10,479
(b) 1995–2000	Northeast		336	1,625	495	2,456
	Midwest	302		1,632	874	2,808
	South	775	1,157		1,097	3,028
	West	357	730	1,281		2,367
	Total	1,434	2,223	4,537	2,465	10,659

where the λ values are simply the natural logarithms of the variables appearing in Eq. (3.1).

In Eq. (3.1), T is the *overall* effect representing the total number of migrants (i.e., n_{++}). This value is "adjusted" (i.e., scaled) by the row and column marginal proportions, O_i and D_j, respectively, leaving the "doubly-constrained" interaction term OD_{ij} as the influence of what Mueser (1989) calls the spatial separation component. More specifically, O_i is the proportion of all migrants leaving from region i (i.e., n_{i+}/n_{++}), D_j is the proportion of all migrants moving to region j (i.e., n_{+j}/n_{++}) and the interaction component OD_{ij} is defined as $n_{ij}/[(T)(O_i)(D_j)]$ or the ratio of the observed number of migrants to the expected number (for the case of no interaction). All effects taken together represent the spatial structure of migration.

The multiplicative components corresponding to the migration flows discussed above are set out in Table 3.2. Note that the overall component (T) is set out in the total sum (i.e., n_{++}) location of the table, the origin components (O_i) are set out in the row-sum locations (i.e., n_{i+}), the destination components (D_j) are set out in the column-sum locations (i.e., n_{+j}), and the origin-destination interaction components (OD_{ij}) are set out in the cells inside the marginal totals (i.e., n_{ij}). For example, consider the 1985–1990 Northeast to South flow of 1,604,000 persons disaggregated into the four multiplicative components:

$$
\begin{aligned}
n_{13} &= (T)(O_1)(D_3)(OD_{13}) \\
&= n_{++}\left(\frac{n_{1+}}{n_{++}}\right)\left(\frac{n_{+3}}{n_{++}}\right)\left[\frac{n_{13}}{(n_{++})\left(\frac{n_{1+}}{n_{++}}\right)\left(\frac{n_{+3}}{n_{++}}\right)}\right] \\
&= (10{,}479)\left(\frac{2{,}403}{10{,}479}\right)\left(\frac{4{,}329}{10{,}479}\right)\left(\frac{1{,}604}{993}\right) \\
&= (10{,}479)(0.2293)(0.4131)(1.6157) \\
&= 1{,}604
\end{aligned}
$$

where the subscripts 1 and 3 denote the Northeast and South regions, respectively. The interpretations of these components are relatively simple. The overall component is the reported total number of U.S.-born interregional migrants aged 5 years and over; 10.5 million persons were counted as interregional migrants between 1985 and 1990. The origin component represents the shares of all migrants from each region: 23% of all migrants originated in the Northeast region. The destination component represents the shares of all migrants to each region: 41% of all migrants moved to the South region. And, finally the interaction component represents the ratio of observed migrants to expected migrants: there were roughly 16 observed migrants for every 10 expected ones. The expected flow is based on the marginal total information, i.e., $(T)(O_1)(D_3)$.

Table 3.2 Saturated log-linear model parameters of U.S. interregional migration flows (in thousands), 1985–1990 and 1995–2000

		Destination				
Period	Origin	Northeast	Midwest	South	West	Total
1985–1990	Northeast	0.0000	0.6635	1.6157	0.8063	0.2293
	Midwest	0.8423	0.0000	1.3662	1.3193	0.2828
	South	1.8018	1.8624	0.0000	1.5127	0.2922
	West	1.2023	1.5514	1.2427	0.0000	0.1957
	Total	0.1378	0.2069	0.4131	0.2422	10,479
1995–2000	Northeast	0.0000	0.6560	1.5542	0.8711	0.2304
	Midwest	0.7996	0.0000	1.3654	1.3457	0.2634
	South	1.9019	1.8320	0.0000	1.5658	0.2841
	West	1.1216	1.4779	1.2707	0.0000	0.2221
	Total	0.1345	0.2085	0.4257	0.2313	10,659

A comparison of these multiplicative components over time for the Northeast to South flow informs us that the overall number of migrants increased by 180,000 (i.e., 10,659,000−10,479,000). The proportions of migrants from the Northeast and to the South both increased, while the interaction term decreased from 1.6 to 1.5. The result was an increase in the number of migrants between these two regions by 21,000 persons.

3.2.4 The "Independence" Model

The saturated log-linear model defined in Eq. (3.2) has reduced forms, also called unsaturated models. The most common of these is the model with no interaction effects. For example, the unsaturated model that only includes the *main effects* of origin and destination is specified as

$$\ln(\hat{n}_{ij}) = \lambda + \lambda_i^O + \lambda_j^D. \tag{3.3}$$

The interregional flows in such a model depend only on origin (row) and destination (column) effects. The model in Eq. (3.3) is often designated (O, D). A model that adds the interaction between origin and destination to Eq. (3.3) would be denoted as (OD). Such notations are used because these models are hierarchical, that is, for two-way interaction terms, the main effect parameters must be included, and for three-way interaction terms (e.g., when age is included) all the main effects and two-way interactions must be included.

What is important to understand about migration flow tables in general, and is illustrated by the independence model, is the importance of the diagonals in the flow matrix (representing the stayers and return migrants). To remove non-migrant elements from the analysis, structural zeros can be inserted using an indicator function (Agresti, 2002; Willekens, 1983a). When structural zeros are *included* in the

Table 3.3 Predicted interregional migration flows (in thousands) under quasi-independence, 1985–1990 and 1995–2000

	Destination				
Origin	Northeast	Midwest	South	West	Total
(a) 1985–1990					
Northeast	0	527	1,314	562	2,403
Midwest	436	0	1,770	757	2,963
South	701	1,142	0	1,219	3,062
West	307	499	1,245	0	2,051
Total	1,444	2,168	4,329	2,538	10,479
(b) 1995–2000					
Northeast	0	527	1,368	561	2,456
Midwest	401	0	1,708	700	2,809
South	690	1,133	0	1,206	3,029
West	343	563	1,462	0	2,368
Total	1,434	2,223	4,538	2,467	10,662

model, Eq. (3.3) is called a *quasi-independence* model. This model predicts migration flows under the condition that origin and destination are independent, and that intra-regional migrations are omitted from the data. When the diagonal elements are replaced in the model by structural zeros, the resulting predicted values under the assumption of independence are much improved. This is illustrated in Table 3.3, which yields R^2 values of 0.86 and 0.88 for the 1985–1990 and 1995–2000 data sets, respectively.

3.3 Biproportional Adjustment and the Method of Offsets

The utility of the multiplicative model extends beyond the convenient decomposition of the observed flow matrix into interpretable parameters that help to describe the spatial structure of migration. The log-linear statistical model also is a powerful instrument for the study of complex data structures. Here, we demonstrate how the log-linear model can be used to predict the migration flows in one period on the basis of flows observed in a previous period. The use of historical data to capture spatial accessibility or spatial interaction hinges on the assumption that spatial interaction effects are stable over time, a hypothesis that has been supported by Willekens (1983a), Nair (1985), Mueser (1989), and Snickars and Weibull (1977), who found that past tables of migrant flows provide much better estimates of current accessibility than any distance measure.

In a number of different applied areas, analysts have used an iterative algorithm to adjust a historical matrix to sum to new row and column marginal totals. Known as the biproportional adjustment method (Bacharach, 1970) or iterative proportional fitting technique, this method effectively imposes the structure found in the historical matrix on the subsequent migration time period.

Consider, for example, the "historical" flow matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

Suppose that the row and column totals are doubled, then clearly

$$\mathbf{B} = \begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix}$$

is a flow matrix with the same interaction effects. What if, instead, only the row totals are doubled and the column totals shift to 4 and 8? How do we impose, as much as possible, the spatial structure of **A** onto the set of marginals? The iterative biproportional adjustment method yields the matrix

$$\mathbf{C} = \begin{bmatrix} 1.123 & 4.877 \\ 2.877 & 3.123 \end{bmatrix}.$$

Notice that the two matrices

$$\mathbf{D} = \begin{bmatrix} 3 & 3 \\ 1 & 5 \end{bmatrix} \quad \mathbf{E} = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$$

also satisfy the marginal constraints, but the spatial structure they exhibit is not biproportional to **A**'s spatial structure.

One may say that the interaction effect measures the preference that a migrant from region i has for region j, if one controls for the differences in the sizes of regional numbers of out-migrants and in-migrants. Migration spatial structure, therefore, has much to do with "destination preference." However, if the destinations that migrants select are independent of the region of origin, i.e., if the probability of selecting a particular destination is the same for all migrants, irrespective of origin, then spatial interaction would be absent, and the flow of migrants from i to j would simply be the product of total out-migrants from i times the probability of selecting destination j (which then is the same for all origins of i). The model, in which the distribution of migrants over destinations is the same irrespective of origin, is known as the *migrant pool model*. The migrant pool model implies the *independence* of origins and destinations.

The contrast between our definition, with its "full" specification of the spatial interaction model (9 parameters for a 2 by 2 flow matrix) and various other definitions is that only with such a detailed specification can one reconstruct the exact flow matrix from the knowledge of its parameters. No other definition proposed thus far does that. The analogy to fully specified model schedules in demography comes to mind. For example, the Heligman-Pollard model mortality schedule, with its eight parameters, can be used to reconstruct quite accurately a whole schedule of age-specific probabilities of dying (Heligman & Pollard, 1980). And the 13-parameter Rogers-Castro model migration schedule can approximate an entire schedule of age-specific probabilities of migrating. Similarly, the nine parameter saturated log-linear model of the above matrix **A** can reconstruct, in this case exactly, the elements of that matrix. Note, most statistical packages use cornered-effect coding, e.g., last category reference coding, for parameter estimation (refer to Raymer, 2007, pp. 989–990). A saturated model using this type of coding scheme would only have as many parameters as there are cells in the table (i.e., 4 parameters). We use the more complicated coding for parameter interpretation and description purposes only. In the actual fitting of the models, we apply the coding scheme adopted by the statistical package being used (e.g., SPSS uses last reference coding).

It turns out that the migrations predicted by the unsaturated log-linear model may also be obtained by the biproportional method. It suffices to replace the interaction term OD_{ij} by a matrix of ones ($OD_{ij} = 1$). The biproportional method is also equivalent to the method of offsets. An *offset*, a matrix with auxiliary information, can be used to incorporate such information (as well as structural zeros) to improve the estimation procedure. Auxiliary information can be, for example, a historical table of migration flows. The log-linear-with-offset model is specified as:

$$\ln\left(\hat{n}_{ij}\right) = \lambda + \lambda_i^O + \lambda_j^D + \ln\left(n_{ij}^*\right) \tag{3.4}$$

where n_{ijx}^* denotes the auxiliary information (refer to Rogers, Willekens et al., 2003, pp. 60–61; Willekens, 1982a, 1983b). In this case the flows contained in the offset would be forced to fit the marginal totals represented by the overall level and main effects of origin and destination.

To illustrate the workings of the method of offsets, consider the log-linear-with-offset model fit of the observed 1995–2000 migration flow matrix in Table 3.1. Suppose we wish to keep the numerical values of the row and column marginal totals, but, at the same time, wish to replace the migration interaction effects observed during that period by the interaction effects observed during the earlier 1985–1990 period, using the method of offsets. What would be the corresponding set of log-linear parameters? Table 3.4 sets out the predicted flow matrix obtained by the method of offsets in Panel (a), and Panel (b) presents the associated multiplicative components. Note that the T, O_i and D_j values of the "predicted" matrix are identical to those found for the observed 1995–2000 flow matrix, but that the other terms (i.e., OD_{ij}) are different, reflecting the changed conditions of the 1995–2000 period.

Finally, Table 3.5 presents the ratios of the two sets of flows: (1) the ratios of the observed 1995–2000 flows structure to that of the observed 1985–1990 flows and (2) the ratios of the predicted 1995–2000 flows (using the 1985–1990 flows as an offset) to the observed 1995–2000 flows. The ratios conveniently indicate the direction of change over the decade: a ratio greater than unity indicates an increased value for the parameter, one less than unity points to a decrease.

Table 3.4 Predicted U.S. 1995–2000 interregional migration flows (in thousands) with observed 1985–1990 as the offset in a log-linear model

	Destination				
Origin	Northeast	Midwest	South	West	Total
(a) Predicted flows					
Northeast	0	323	1,667	466	2,456
Midwest	312	0	1,619	877	2,808
South	744	1,163	0	1,123	3,030
West	378	737	1,252	0	2,367
Total	1,434	2,223	4,538	2,466	10,661
(b) Multiplicative components					
Northeast	0.0000	0.6307	1.5946	0.8203	0.2304
Midwest	0.8260	0.0000	1.3545	1.3502	0.2634
South	1.8255	1.8408	0.0000	1.6023	0.2842
West	1.1873	1.4932	1.2426	0.0000	0.2220
Total	0.1345	0.2085	0.4257	0.2313	10,661

Table 3.5 U.S. interregional migration: Ratios of observed 1995–2000 flows to observed 1985–1990 flows and predicted 1995–2000 flows

	Destination				
Origin	Northeast	Midwest	South	West	Total
(a) Observed 1995–2000/Observed 1985–1990					
Northeast		1.0199	1.0133	1.0545	1.0222
Midwest	0.8781		0.9757	0.9227	0.9474
South	1.0190	0.9803		0.9772	0.9888
West	1.0513	1.1083	1.2163		1.1543
Total	0.9930	1.0252	1.0481	0.9712	1.0172
(b) Predicted 1995–2000*/Observed 1995–2000					
Northeast		0.960	1.026	0.942	1.000
Midwest	1.033		0.992	1.004	1.000
South	0.960	1.005		1.024	1.001
West	1.058	1.010	0.978		1.000
Total	1.000	1.000	1.000	1.000	1.000

*Predicted 1995–2000 represents the 1995–2000 main effects model with 1985–1990 as offset

3.4 Introducing Additional Structures

3.4.1 Overview

The multiplicative component and log-linear models described in the previous section can be readily extended to include other categorical variables, such as age, sex, ethnicity, health status, economic activity and so on. In this section, we illustrate an extension that includes *age-specific* migrant patterns between regions. The multiplicative component model for this table is specified as:

$$n_{ijx} = (T)(O_i)(D_j)(A_x)(OD_{ij})(OA_{ix})(DA_{jx})(ODA_{ijx}) \tag{3.5}$$

where A_x is the proportion of all migrants in age group x. This model is more complicated because there are now three two-way interaction components and a single three-way interaction component between the origin, destination, and age variables. However, the interpretations of the parameters remain relatively simple and follow the same format as presented for the two-way table. That is, the interaction components represent ratios of observed flows or marginal totals to expected ones. For example, the destination-age interaction (DA_{jx}) component is calculated as $n_{+jx}/[(T)(D_j)(A_x)]$ and represents the ratio of observed to expected for in-migrants of age x to region j.

Unsaturated log-linear models of the (saturated) multiplicative model set out in Eq. (3.5) are useful for understanding the importance of age and its interaction with origin and destination. For example, the main effects log-linear model is specified as

$$\ln(\hat{n}_{ijx}) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A \tag{3.6}$$

This model assumes independence between each of the categories of origin, destination, and age and is designated (O, D, A). A model that includes the interaction between origin and destination plus all of the main effects is designated as (OD, A) with its corresponding model specification being:

$$\ln(\hat{n}_{ijx}) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD}. \tag{3.7}$$

Auxiliary information also can be incorporated. For example, the model

$$\ln(\hat{n}_{ijx}) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \ln\left(n_{ijx}^*\right) \tag{3.8}$$

forces the values in the offset to fit the marginal totals represented by the overall level and main effects of age, origin, and destination. The two-way and three-way association structures, i.e., λ_{ij}^{OD}, λ_{ix}^{OA}, λ_{jx}^{DA} and λ_{ijx}^{ODA}, contained in the offset, however, remain the same.

3.4.2 Descriptive Analysis

We continue our analysis of migration between the four regions in the United States during the 1995–2000 using the multiplicative components model set out above. Such an analysis follows a hierarchical format, starting with the overall level component and ending with the three two-way interaction components. The three-way interactions between origin, destination, and age are not analyzed for two reasons.

The first is that most of the structure found in the migration patterns is captured by the overall, main, and two-way interaction effects. The second reason is that, although patterns are often found in the three-way interactions, it is tedious to incorporate them into the modeling process, and their interpretation is more difficult. Therefore, we shall just focus on the simpler and more powerful aspects of the model represented by the other seven terms found in Eq. (3.5).

The extension to include age is straightforward. The age groups used in this chapter start with 5–9 years and end with 85+ years and are measured at the time of the census. There are seventeen age groups in total. As illustrated in Fig. 3.2, the age main effect components for the 1985–1990 and 1995–2000 periods describe the age composition of all migrants in the two multiregional systems. The origin-age interaction components can be used to identify important differences between age-specific out-migration from each region and the overall age profile of migration found in the corresponding expected flows (i.e., $(T)(O_i)(A_x)$). The same is true for the destination-age interaction components, but with a focus on the differences between age-specific numbers of in-migrants to each region and their corresponding expected flows (i.e., $(T)(D_j)(A_x)$).

The origin-age and destination-age interaction components are useful for identifying the relative differences found in age patterns of in-migration and out-migration, respectively. For example, in examining the origin-age components set out in Fig. 3.3, we find higher propensities of young adult and elderly migration from the Northeast and Midwest regions. The opposite was true for those from the South and West regions. The patterns over time show that, for example, the relative numbers of elderly migration from the Northeast were lower in 1995–2000 than they were in 1985–1990. The destination-age interaction components displayed in

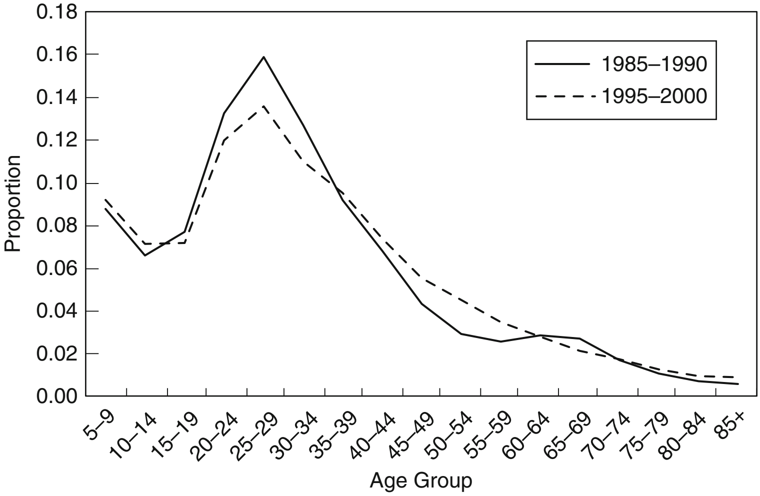


Fig. 3.2 The age main effect component (A_x): U.S. interregional migration, 1985–1990 and 1995–2000

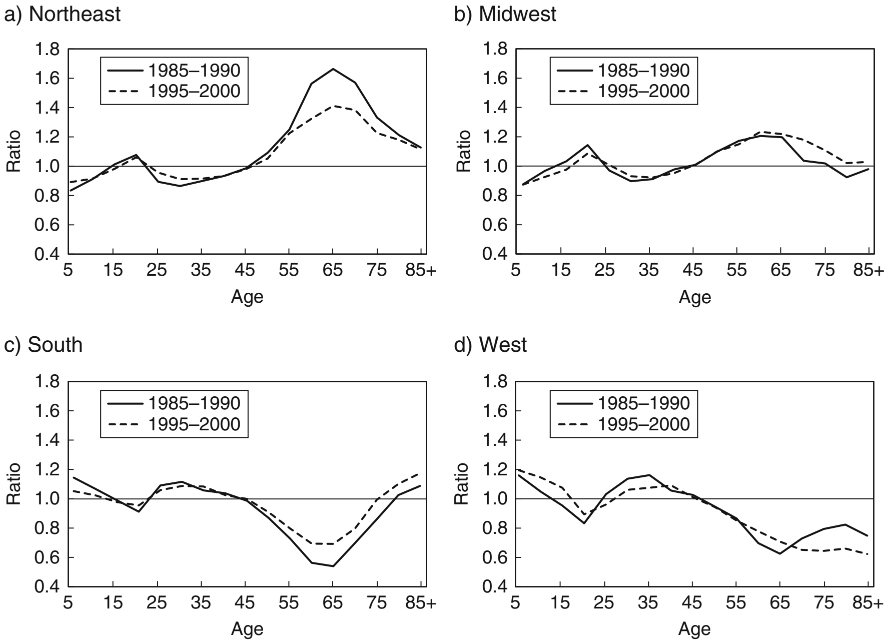


Fig. 3.3 The origin-age interaction components (OA_{ix}): U.S. interregional migration, 1985–1990 and 1995–2000

Fig. 3.4 generally exhibit the opposite patterns of those found in the origin-age components. The main exceptions are the migrations of the elderly from and to the West, both of which show lower than expected patterns.

3.4.3 Unsaturated Log-Linear Model Analysis

In this section, we compare different *unsaturated* log-linear models to analyze underlying structures in the 1995–2000 U.S. migration data. All models include structural zeros to remove non-migrants from the predictions and the estimations. The results are set out in Table 3.6, and the models are compared using the likelihood ratio statistic, G^2,

$$G^2 = 2\sum n_{ijx} \ln(n_{ijx}/\hat{n}_{ijx}), \tag{3.9}$$

where $\hat{n}_{ijx}$ denotes the predicted age-specific migration flows, and values of G^2 closest to zero are associated with "good" fits (see, e.g., Agresti, 2002).

The most obvious finding is that the origin-destination interaction term is very important for accurately predicting the age-specific migration flows. Most of the flows do not contain a large retirement peak or major deviations from the overall

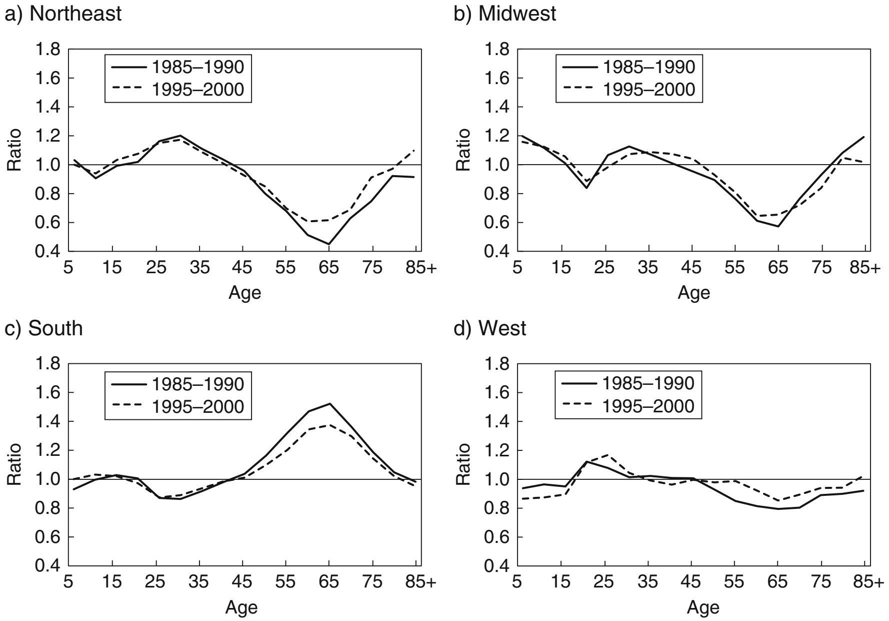


Fig. 3.4 The destination-age interaction components (DA_{jx}): U.S. interregional migration, 1985–1990 and 1995–2000

Table 3.6 Unsaturated log-linear model fits: Age-specific interregional migration flows in the U.S., 1995–2000

Model	Likelihood ratio statistic, G^2	Residual degrees of freedom, df	G^2/df
(ODA)	0	204	0
(O, D, A)	573,446	181	3,168
(OD, A)	281,056	176	1,597
(OA, D)	433,718	133	3,261
(DA, O)	414,899	133	3,120
(OD, OA)	141,328	128	1,104
(OD, DA)	122,509	128	957
(OA, DA)	321,318	85	3,780
(OD, OA, DA)	43,875	80	548

Note: Residual degrees of freedom = number of non-redundant parameters in saturated model (ODA) minus the number of parameters in an unsaturated model

age profile of migration. However, the fits are slightly improved when the origin-age or destination-age interactions (with the latter doing a better job) are included. Of course, to capture different age profiles found in some of the flows, such as those with retirement peaks, origin-age or destination-age interactions have to be included. Figure 3.5 provides an example of the *(O, D, A)*, *(OD, A)* and *(OD, OA,*

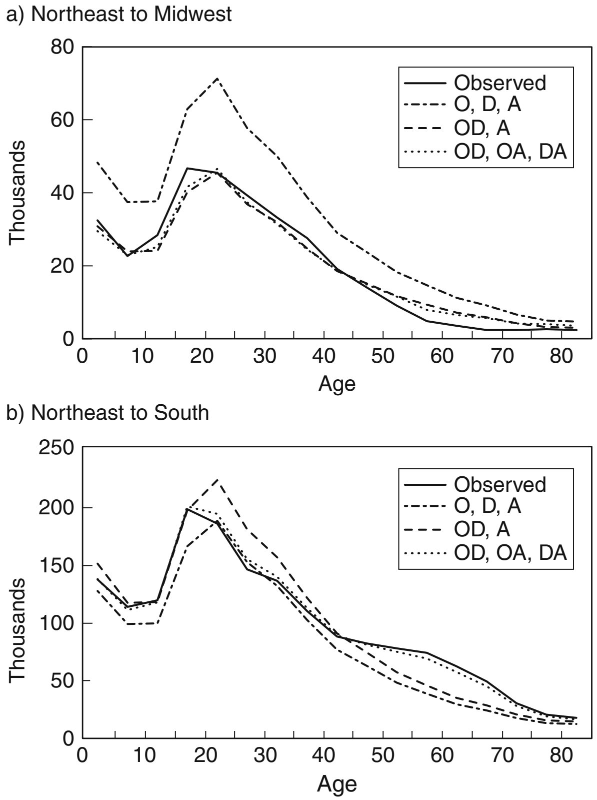


Fig. 3.5 A comparison of unsaturated log-linear model fits: Northeast to Midwest and South, 1995–2000

DA) fits in relation to the observed values for the Northeast to Midwest and South flows during the 1995–1990 periods.

3.5 Summary and Discussion

What do we mean when we refer to the spatial structure of migration? This expression has been used rather loosely in the literature and needs to be defined more rigorously if it is to be of much use as a tool for comparative analysis of flows or for developing indirect methods of estimating migration streams in the absence of flow data. One way to define migration spatial structure is to draw on the demographer's way of defining age structure, i.e., as the proportional distribution of the numbers of persons enumerated at each age or in each age group. Thus if one were to double the total population, but leave the proportional distribution unchanged, one would conclude that the population increased, but that its age structure remained unchanged. Adapting this definition for the migration structure of a region's destination-specific out-migration streams, one could define that structure to mean the proportional distribution of the total outflow across the set of alternative destinations. In that case, if a doubling of the region's out-migration level were distributed in the same proportional manner over age groups and destinations, then one would conclude that the migration spatial structure had remained the same as before. This definition, however, only makes sense in a linear model of the phenomenon. If one instead adopts a non-linear "gravity model" type of formulation—say a spatial interaction model representation of origin-destination-specific migration flows—then clearly one also needs to consider the change of the destination population and also the separation effect between each origin-destination pair of locations. Thus the impact on spatial structure of a doubling of the migration outflow needs to be considered in tandem with the change in the destination population size or size of migration inflow. For example, the impact of a tripling would be different than that of a doubling. *What this implies then is that a full specification of the spatial interaction model needs to be used in the definition of migration spatial structure.* If this is true, then how do we interpret the use of a historical matrix to predict current migration? We interpret it as the migration spatial structure we wish to impose on a current set of marginal totals.

The various indicators of migration spatial patterns that have been popular in the literature describe only particular attributes of a particular migration spatial structure: for example, its efficiency in redistributing the multiregional population, or its spatial focus, or, indeed, its implicit destination preferences. None of these could be used to impose a unique historical migration spatial structure onto a current situation. They allow only a partial assessment of comparative structures, and they, therefore, are of limited use as tools of indirect estimation. However, as partial indicators of different attributes of spatial patterns, they can and have played a useful role in comparative studies of such patterns. The relevant literature is rich with examples of the useful findings generated by indices of migration efficiency and of spatial focus, for example. However, we believe the log-linear model introduces the influence of the separation (or interaction) effects more fully, and it also seems to bring in the relative population size effects more directly.

Chapter 4
Smoothing Age and Spatial Patterns

4.1 Introduction

A comparison of an observed pattern of age-specific rates or probabilities with the corresponding model schedule fitted pattern identifies idiosyncrasies in the observed data and points to possible data errors or to irregularities created by an insufficiently large sample. Actuaries calculating life insurance policies or annuities, for example, would want to smooth irregular patterns to ensure that age-specific probabilities of dying, do not show, say, that an average 45-year old female had a higher risk of dying within the next year than did an average 46-year old female. Confronting such an irregularity, an actuary is likely to *smooth* out the suspicious behavior with a model mortality schedule, for example, the eight-parameter Heligman-Pollard (1980) model mortality schedule.

An analogous problem is illustrated in Fig. 2.4 in Chapter 2 by the observed out-migration schedule in 1974 for Stockholm males moving to the rest of Sweden. In this particular illustration ($x + A$)-year old males exhibit a higher out-migration rate than do men a year younger. This pattern is suspicious because in most "normal" migration schedules one finds a monotonic decrease in rates for males in their late twenties and early thirties. Thus demographers may wish to *smooth* out such suspicious behavior with a model migration schedule, such as the overlaid Rogers-Castro schedule that also appears in Fig. 2.4.

What do we mean by *smoothing*? In this book we follow the definition published in the United Nations (1983) manual on indirect estimation:

> The term "to smooth" is used in this *Manual* in its most general sense to mean elimination or minimization of irregularities often present in reported data. (United Nations, 1983, p. 147 fn.)

Although migration patterns normally are thought to change in smooth and gradual increments across the life span, observed data-based patterns are often jagged and irregular. This can be attributed to the random variation that inevitably accompanies survey data and from the aggregation of data into convenient intervals for reporting purposes. Three smoothing techniques are presented in this chapter, each of which is designed to reduce the effects of randomness and aggregation. These

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_4,

smoothing procedures are: (1) using splines to interpolate migration data that are based on 5-year age groupings into single-year migration age profiles (McNeil et al., 1977), (2) fitting model migration schedules to the splined age profiles (Rogers and Castro, 1986), and (3) using log-linear models to simplify irregular migration patterns observed in a contingency flow table (Raymer & Rogers, 2007).

We begin in Section 4.2 to apply smoothing methods to the full sample age-specific state-level migration data, which are reported by the U.S. Census Bureau in 5-year age aggregations. These data provide migrant age profiles that are crude step functions (histograms), needing to have the jagged steps transformed into smooth and continuous profiles, often by the application of splines. For even more refined smoothing, the splined profiles then may be fitted with model migration schedules. Because the initial data used in this section are based on large sample estimates of migration patterns, the refinements introduced by the smoothing procedures result in qualitatively improved migration profiles that are thought to be more true to the unobserved patterns in the population.

Section 4.3 provides a demonstration of how smoothing techniques can improve the accuracy as well as the regularity of migration age profiles. This demonstration is also based on data drawn from the Census 2000 full sample data, which was made available for analyses of individual observations in the Census 2000 Public-Use Microdata Sample (PUMS) 1% data files. After smoothing procedures are applied to age-specific migration estimates, based on this substantially reduced sample, they produce age profiles that conform more closely to the profiles derived from the full sample data.

The results of Section 4.3 suggest that the smoothing methods do produce improvements in the reliability of the migration age profiles that are derived from the less reliable Census 2000 PUMS 1% sample data. In Section 4.4 we use the smoothing methods developed in the previous sections and apply them to data derived from the 2005, 2006, and 2007 American Community Survey (ACS) PUMS samples for U.S. states. These data are substantively different from the migration data of prior decennial censuses. In addition, the ACS provides yearly estimates of annual migration age patterns, which are shown to be even less reliable than the 5-year migration age patterns estimated from the Census 2000 PUMS 1% sample data. Nevertheless, applying the smoothing methods to the ACS-based estimates demonstrates that improvements are thereby gained in the regularity and the reliability of migration age profiles.

Section 4.5 introduces a log-linear specification of the smoothing problem, and Section 4.6 offers a summary and a concluding discussion.

4.2 Smoothing Irregular Migration Data: Census 2000 Full Sample

The Census 2000 Long-Form Survey was a 16.67% national probability sample. It was the primary source of U.S. migration data until the full implementation of the ACS in 2005. Due to its large sample properties, the Census 2000 full sample estimates of state-level migration are very accurate, and yet they still show irregularities caused by randomness and the aggregation methods used to prepare the published tables. Smoothing operations reduce these irregularities at the same time that they transform the step functions, which reflect the migration patterns reported in broad age intervals, into smooth and continuous age profiles. After smoothing the full sample data, we define the resulting migration age profiles as standards, which are compared to the reduced sample estimates obtained from the PUMS 1% sample. The results of these comparisons serve as the basis for assessing the accuracy of the small sample estimates, as well as for assessing the improvements in accuracy that result from the smoothing procedures.

The full sample age-specific out-migration data used in this chapter were aggregated from the Census 2000 Long-Form Survey for the fifty states and the District of Columbia. They were distributed on the Census 2000 Migration DVD released by the U.S. Census Bureau, reporting counts of persons who left their state of residence between 1995 and 2000 and lived to be counted as residents of another state by the 2000 census. Based on a person's age in 2000, these counts were tabulated into 5-year age categories, beginning at age 5 and ending at age 85 or older, i.e., ages 5–9, 10–14, 15–19,. . ..80–84, 85+. Using these data, we backcasted persons to where they lived five years earlier in order to associate out-migrants with their respective state of origin and age category in 1995. For example, those persons aged 5–9 in 2000 became aged 0–4 in 1995, and those aged 20–24 in 2000 became aged 15–19 in 1995. After the backcasting, the 1995 age categories range from ages 0–4 to ages 80+, in contrast to the age categories of migrants reported in 2000, which range from ages 5–9 to ages 85+.

The procedure for counting the numbers of persons at risk for migrating from each state and for each age grouping is similar to the procedure used for counting the numbers of out-migrants. The persons counted by the U.S. Census Bureau as living in the U.S. in 2000 were assigned to their 1995 state of residence and to their 5-year age category in 1995. The migration propensity is the number of persons in a particular age category who migrated out of state between 1995 and 2000 divided by the total number of persons in the age category who were living in the state in 1995. Because this proportion is conditioned on a person's survival to the year 2000, it is sometimes called a conditional survivorship and denoted $S_i(x, x+4)$ or $S_i(x)$, where i represents the state of residence in 1995 and (x, x+4) denotes a specific 5-year age category. Panel (a) of Fig. 4.1 shows the profile of migration propensities that result for the state of Indiana.

To arrive at smooth age migration profiles, the initial migration proportions for the 5-year age categories are assigned values close to the middle age within the 5-year interval, i.e., ages 2, 7, 12, 15, . . .72, 77. From this set of points a continuous age profile of state out-migration propensities is generated with cubic spline interpolation, which constructs third-order polynomials that pass through the set of pre-defined control points. Cubic spline interpolation provides a smooth profile for all integer values of ages between 0 and 79, using 2, 7, 12, . . .,77 as the nodes for the spline algorithm as implemented by Advanced Systems and Design, which is as an add-on function for Microsoft Excel. The splined results for Indiana are displayed in Panel (b) of Fig. 4.1.

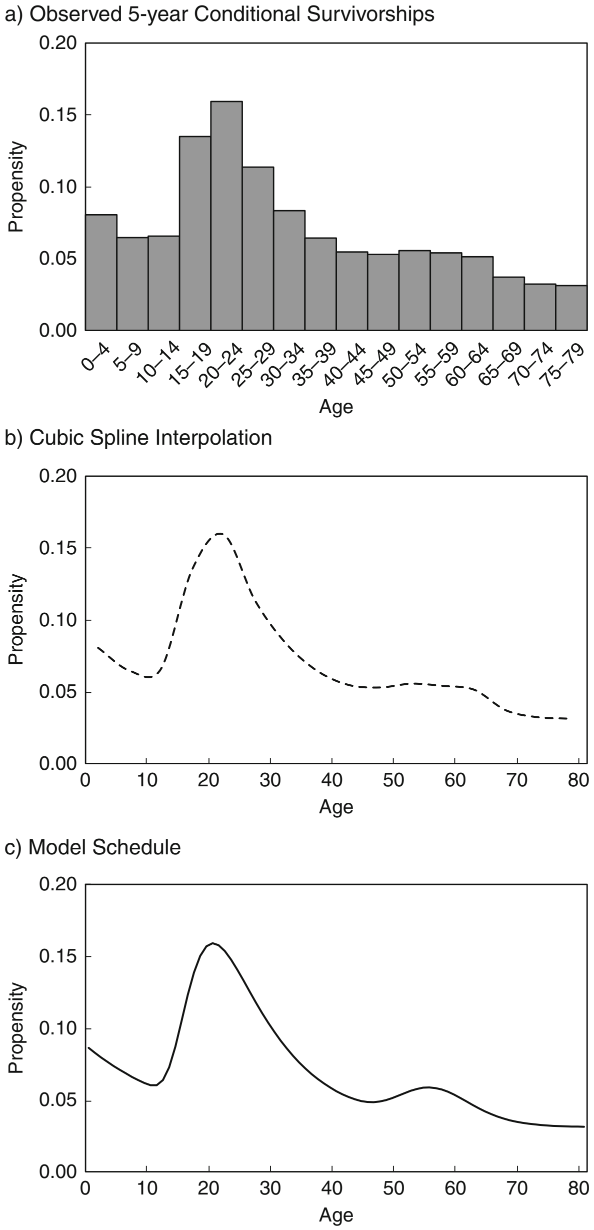


Fig. 4.1 Census 2000 full sample migration propensities: A demonstration of smoothing procedures for Indiana

The cubic-splined data set is then fitted by the appropriate Rogers-Castro model migration schedule using a nonlinear regression program developed in MATLAB, producing thereby a final set of 1-year age propensities and a smooth curve that adheres to the known regularities of migration age profiles and, at the same time, preserves the observed levels of migration. Panel (c) of Fig. 4.1 illustrates the contrast of the model migration curve to the splined profile in Panel (b). The model schedule fit removes the irregularities present in the cubic spline interpolation.

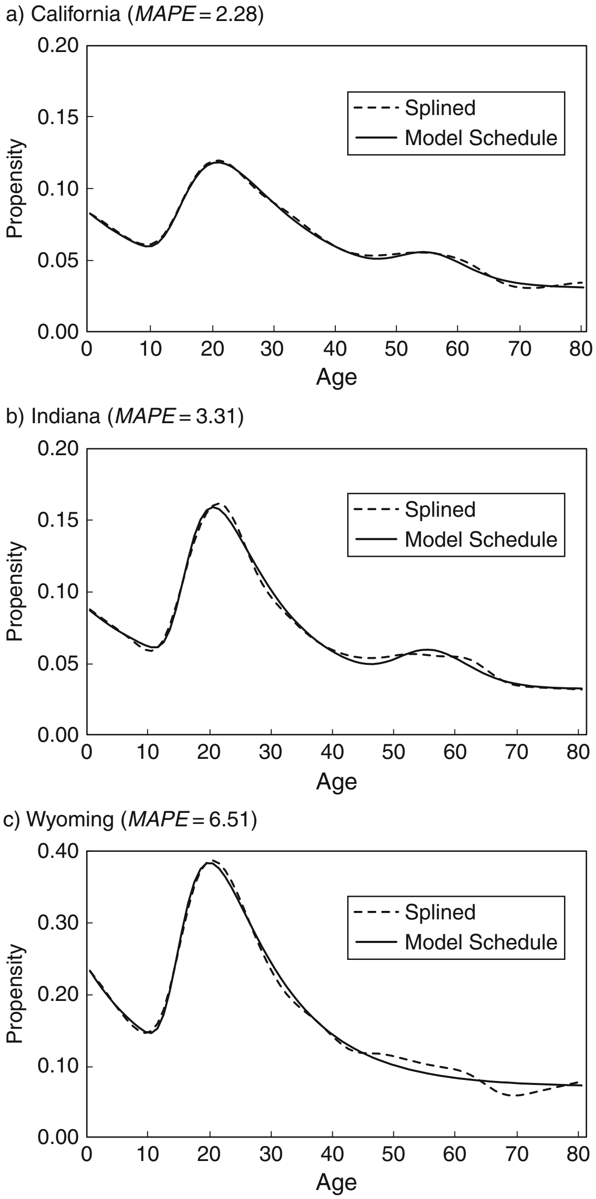
Figure 4.2 shows the new profiles that resulted from the splining process and from the model schedules that were fitted to the splined data for three demonstration states. The splined profiles are smooth; however, there are some irregularities that are smoothed further by the model schedule fits. The closeness of the fits between the model schedules and the cubic splines is measured by the Mean Absolute Percent Error (*MAPE*) statistic:

$$MAPE = 100^{*}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{|F_i - O_i|}{O_i}\right] \tag{4.1}$$

where the absolute value of the difference between the fitted model schedule value (F_i) and the corresponding splined value (O_i) is divided by the splined value (O_i), for each age i. These quotients are averaged over all ages n and multiplied by 100 to arrive at the mean absolute percentage error.

The MAPE scores vary across all fifty states and the District of Columbia, ranging from 2.28 in California to a 10.26 in Utah. The average *MAPE* is 5.56. The slight differences between the two profiles, even for the least populated states, is evidence of the accuracy of the initial full sample data, and it supports our strategy of accepting these model schedules as the best representation of the “true” migration age structures for all 50 states and the District of Columbia.

Further justification for using the model schedules, derived from the full sample data, to represent the true migration schedules was gained by calculating the averaged 5-year propensities (from the model schedules) and by comparing them to the corresponding actual full sample data as released by the U.S. Census Bureau. The outputs of this exercise are illustrated in Fig. 4.3, which displays the results for the same three demonstration states described in Fig. 4.2. The “Observed” refers to the propensities derived from the tabulated data reported by the U.S. Census Bureau, and the “Model Schedule (5-Year Average)” refers to the propensities obtained from the model schedule values, summed and averaged over each of the 5-year age groupings. For the three selected states, the *MAPE*s are 2.62 for California, 3.22 for Indiana, and 7.17 for Wyoming, which are quite small, even for the least populated state of Wyoming.

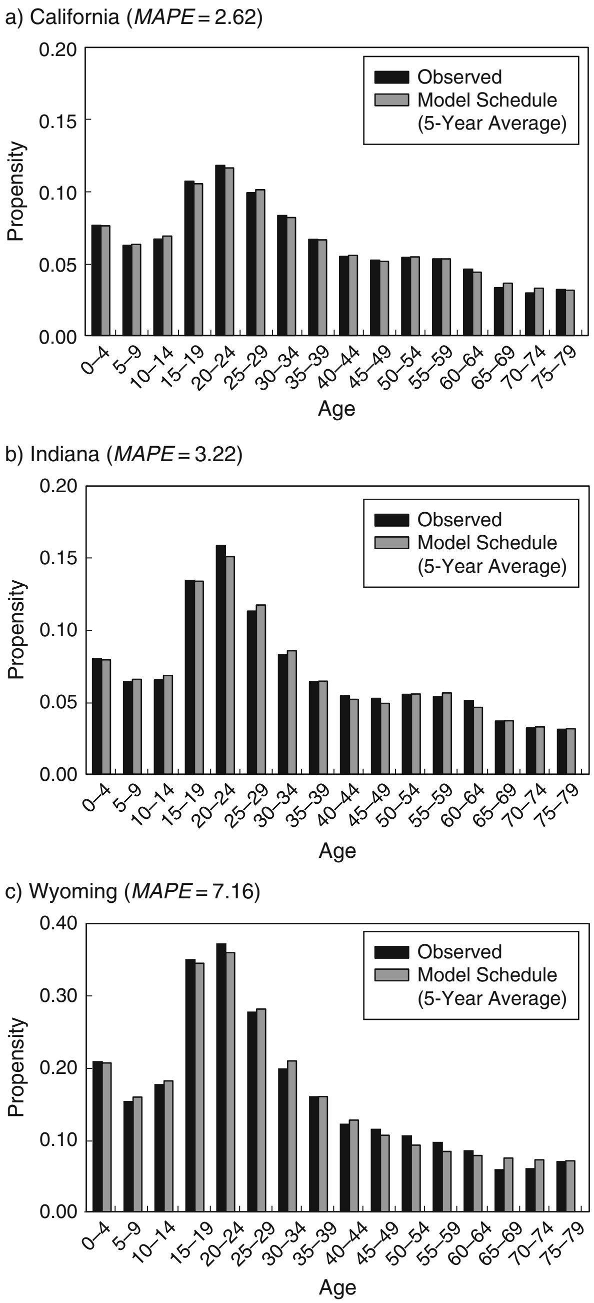


a) California (*MAPE* = 2.28)

b) Indiana (*MAPE* = 3.31)

c) Wyoming (*MAPE* = 6.51)

Fig. 4.2 Migration age patterns of state out-migrants derived from Census 2000 full sample data: Splined patterns compared to model schedules


a) California (*MAPE* = 2.62)

b) Indiana (*MAPE* = 3.22)

c) Wyoming (*MAPE* = 7.16)

Fig. 4.3 Migration age patterns of state out-migrants derived from Census 2000 full sample data: Observed propensities and model schedules (averaged over 5-year age intervals)

4.3 Smoothing Irregular and Inadequate Migration Data: Census 2000 PUMS 1% Sample

The Census 2000 full sample data provide estimates of migration age patterns, which in general are somewhat irregular. However, the underlying robustness of the full sample estimates allow them to be transformed into smooth and reliable migration age structures that are deemed to be accurate representations of state out-migrating populations. But demographers rarely have the luxury of working with such large sample surveys. It has become more typical to have access to PUMS data such as the Census 2000 PUMS 1% sample data, and the annual ACS PUMS samples, which also total approximately 1% of the U.S. population.

The Census 2000 PUMS 1% sample is a substantially reduced sample drawn from the Census 2000 full sample, and the contrast between the two samples offers a unique opportunity to gauge the effectiveness of the smoothing techniques applied to survey data that are roughly comparable in sample size to the ACS PUMS samples. By contrasting the Census 2000 PUMS 1% sample data with the migration schedules that resulted from smoothing the full sample data, we demonstrate the initial problems with the 1% sample estimates, and, in addition, demonstrate the improvements that are gained by applying the smoothing techniques to the smaller sample estimates. The strategy of comparing the migration age patterns based on the data from a smaller sample with those based on the full sample survey allows us to make judgments about the likely improvements that can be gained by applying these same smoothing techniques to survey data such as the ACS PUMS.

The Census 2000 PUMS 1% sample data were obtained from the IPUMS USA website http://usa.ipums.org/usa/. The first step of processing selects out the persons who moved from one state to another between 1995 and 2000. These are the persons who have a state of residence in 2000 that is different from their state of residence in 1995. The PUMS 1% sample data file includes a variable which specifies each person's age on April 1, 2000. Because the migration question asked where the person was living five years ago, the age of a migrant in 2000 is necessarily at least five, since anyone younger was not alive in 1995. The second processing step assigns the age in 1995 to be 5 years younger than the age in 2000. So the age distribution of persons in 1995 goes from 0, 1, 2, . . ., 90 and higher. The individual observations in the data are weighted to inflate the sample to reflect the estimated size of the population in each respective age category within each state of residence in 1995.

The initial age propensities were calculated using the numbers of out-migrants in each single-year age category, divided by the total state population in each single-year age category in 1995. These profiles are very irregular due to the instability of the small sample estimates for single-year age groupings. Panel (a) of Fig. 4.4 demonstrates the sporadic nature of the observed out-migration profile for New Hampshire, one of the least populated states in the nation. The migration propensities obtained from the small sample data, when compared to the model schedule derived from the full sample data, yielded a *MAPE* equal to 31.53.

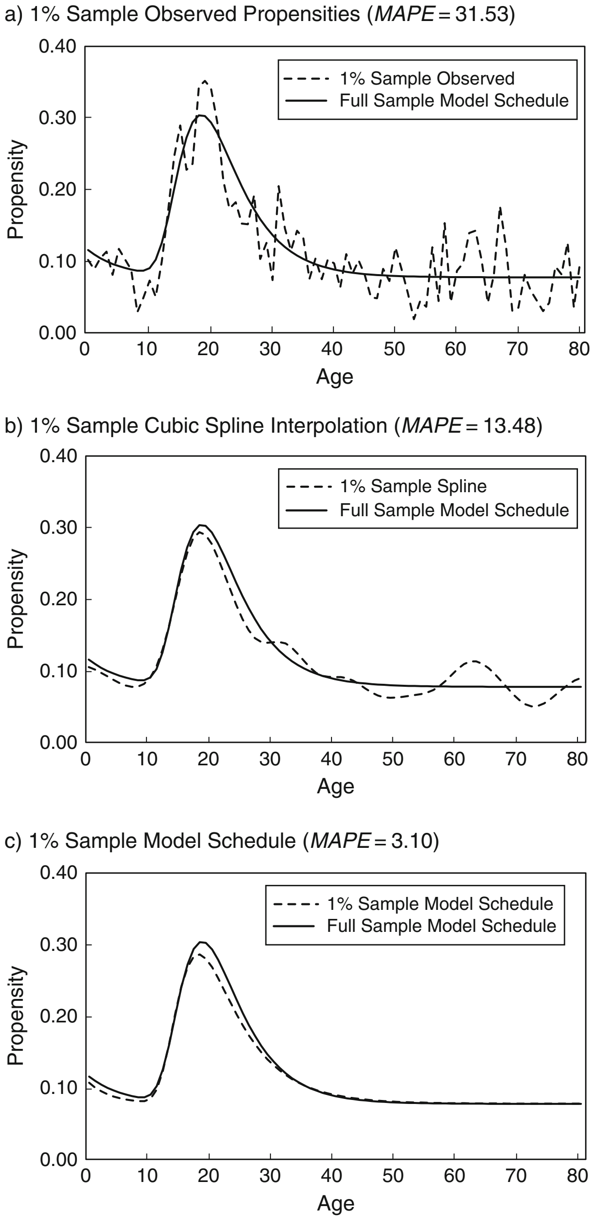


Fig. 4.4 The observed migration propensities from the Census 2000 PUMS 1% sample compared to the model schedule fits of the Census 2000 full sample: A demonstration of the smoothing procedures for New Hampshire

Since the Census 2000 full sample data are reported by the U.S. Census Bureau in 5-year age aggregations, the Census 2000 PUMS 1% sample state out-migrants also were aggregated from the numbers of migrants in each single-year age category into 5-year age categories, making them more comparable to the full sample data and reducing the instability of the single-year estimates. After aggregating the single-year of age out-migration numbers, the smoothing procedures for the small sample data were identical to those used on the full sample data. In other words, the numbers in each 5-year age category were distributed evenly across the single-year ages, and a cubic spline interpolation yielded a smooth profile for all integer values of ages between 0 and 79, using 2, 7, 12, . . .,77 as the nodes for the spline algorithm. Panel (b) in Fig. 4.4 demonstrates the resulting splined profile for the State of New Hampshire, and it clearly shows that the splining procedure yielded improved regularity as well as improved accuracy. The *MAPE* was reduced from 31.53, associated with the observed propensities, to 13.48 for the splined profile.

To complete the smoothing operations on the Census 2000 PUMS 1% sample data, and to make them more comparable to those used on the full sample, the splined profile was fitted with the 7-parameter Rogers-Castro model migration schedule. The result of this step is shown in Panel (c) of Fig. 4.4, and the *MAPE* was reduced further from 13.48 for the splined profile to 3.10 for the smaller sample model schedule as compared to the full sample model schedule.

In summary, our smoothing of the age propensities of migration produced by a small sample survey is a three-step process. This is demonstrated in Fig. 4.4, where each panel shows the profile that resulted from one step of the smoothing process. Ultimately, the degree of correspondence between the model schedule of the small sample data and the model schedule of the full sample data is visually striking. At each step, the estimates of the migration propensities, were improved by the smoothing procedure.

Table 4.1 shows the incremental improvements in reliability for each of the fifty states and the District of Columbia. The column means are reported at the bottom of Table 4.1 and they show that, on average, for all states and the District of Columbia, the error was reduced from the average $MAPE = 21.11$ associated with the first step of the smoothing process to the average $MAPE = 5.61$ associated with the final step. It is clear from an inspection of Table 4.1, that the errors are most dramatic for the single-age migration profiles, and these errors can be reduced substantially through aggregation into 5-year age categories followed by the cubic spline interpolation procedure. Nebraska and Wyoming are two additional examples of less populated states that initially showed substantial errors, with *MAPE*s = 24.8 and 35.42, respectively, which then were reduced to 14.57 and 8.00, respectively, by the splining process. Fitting model schedules to the splined profiles offered additional improvements in accuracy for most of the states. For example, the Nebraska and Wyoming model schedules reduced the error from 14.57 to 8.00 and 11.79 to 2.73, respectively.

In general, the smoothing procedures seem to improve the reliability of state migration schedules derived from survey data, regardless of the state population sizes. However, the amount of improvement is clearly related to sample size.

Table 4.1 The *MAPE*s associated with the observed, the splined, and the model schedules derived from the Census 2000 PUMS 1% sample data

	Observed	Splined	Model schedule
1995 State of origin	*MAPE*	*MAPE*	*MAPE*
Alabama	20.61	8.59	2.94
Alaska	36.84	18.06	10.40
Arizona	13.35	5.59	1.83
Arkansas	20.81	7.90	4.39
California	6.65	3.91	2.54
Colorado	16.19	8.19	4.73
Connecticut	21.28	11.59	7.69
Delaware	33.12	11.99	6.20
District of Columbia	30.88	19.49	13.51
Florida	9.53	6.52	1.93
Georgia	14.49	6.26	3.44
Hawaii	27.76	14.63	9.32
Idaho	27.94	14.15	4.57
Illinois	11.30	5.72	4.27
Indiana	16.11	6.40	4.66
Iowa	22.21	12.74	3.71
Kansas	18.87	13.63	6.02
Kentucky	22.46	14.11	6.16
Louisiana	17.93	11.82	3.78
Maine	34.53	17.58	5.30
Maryland	15.67	10.29	4.34
Massachusetts	16.96	8.15	7.05
Michigan	14.09	7.59	6.16
Minnesota	20.41	9.81	8.29
Mississippi	26.36	13.60	7.67
Missouri	16.65	8.78	2.77
Montana	32.92	17.17	9.03
Nebraska	24.86	14.57	11.79
Nevada	19.04	9.36	4.57
New Hampshire	31.53	13.48	3.10
New Jersey	9.78	3.86	2.60
New Mexico	22.99	7.07	2.52
New York	7.52	4.14	1.61
North Carolina	14.17	6.92	1.48
North Dakota	35.53	18.66	10.18
Ohio	14.75	10.74	7.18
Oklahoma	19.37	8.88	6.03
Oregon	16.18	6.59	3.33
Pennsylvania	13.69	8.90	2.24
Rhode Island	36.61	15.81	4.66
South Carolina	18.70	8.03	5.71
South Dakota	32.14	15.43	12.55
Tennessee	16.15	4.85	2.81
Texas	9.75	6.34	3.00
Utah	25.36	11.62	7.93

Table 4.1 (continued)

	Observed	Splined	Model schedule
1995 State of origin	*MAPE*	*MAPE*	*MAPE*
Vermont	40.63	18.73	13.33
Virginia	14.74	8.44	4.12
Washington	14.30	7.21	5.31
West Virginia	22.60	14.37	10.63
Wisconsin	21.91	12.61	0.94
Wyoming	35.42	8.00	2.73

Table 4.2 The average *MAPEs* associated with the observed, splined, and model schedules derived from the Census 2000 PUMS 1% sample data by categories of state population size (decreasing)

		Observed average	Splined average	Model schedule average
State population (in millions)	N	*MAPE*	*MAPE*	*MAPE*
16 or more	3	7.98	4.80	2.38
15.99−11.00	3	12.32	7.97	3.90
10.99−6.00	5	13.90	6.55	4.15
5.99−5.00	6	16.80	8.36	3.47
4.99−4.00	5	17.41	8.85	4.19
3.99−3.00	6	19.03	9.56	5.61
2.99−2.00	5	22.72	11.90	5.94
1.99−1.00	9	27.54	13.45	6.27
0.99−0	8	34.68	15.94	9.74

Table 4.2 contrasts the accuracy of the migration profiles derived from the Census 2000 PUMS 1% sample data as compared to the full sample model schedules with respect to the number of sampling units in the survey, or more precisely, the size of the state population. Here we use the state's population in 1995 as backcasted from the Census 2000 PUMS 1% sample data. The average *MAPE*s are reported in Table 4.2 by categories of population size (in decreasing order). For the 26 states with populations of over four million, the average *MAPE* is 14.37 when the observed profiles, derived from the Census 2000 PUMS 1% sample data, are compared with the full sample model schedules. For the 25 states (including the District of Columbia) with populations under four million, the comparable average *MAPE* is 26.90. The disparity in reliability between the more and the less populated states is diminished with the application of model schedules. For the more populated states (more than 4 million) the average *MAPE* was reduced from 14.37 (for the observed profiles) to 3.71 (for the model schedules). For the less populated states (less than 4 million) the average *MAPE* was reduced from 26.90 (for the observed profiles) to 7.06 (for the model schedules).

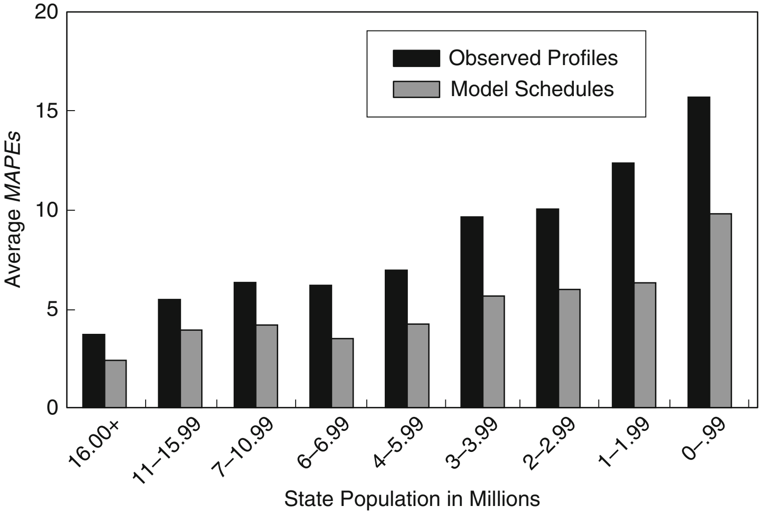


Fig. 4.5 A comparison of the observed profiles and the model schedules derived from the Census 2000 PUMS 1% sample with the model schedule fits of the Census 2000 full sample, with respect to state population size

The relationship between population size and an increase in reliability resulting from smoothing procedures is illustrated in Fig. 4.5, which shows that the smoothing procedures yield accurate profiles of the age propensities for state out-migration, especially in the most populated states. The smoothing procedures consistently improved the accuracy of the small sample based profiles, but the average *MAPE*s were still larger for the less populated states than for the more populated states. These results suggest that, on average, errors are larger when associated with smaller sample sizes, even after the smoothing procedures. However, the percentage of reduction in error, due to smoothing, was about the same for the states with populations less than 4 million as compared to the states with populations more than 4 million (74%). In other words, the model schedules that result from the smoothing procedures may not be as reliable for the less populated areas as for the more populated areas, on average, but the proportionate reduction in error will likely be roughly constant for sample sizes similar to the Census 2000 PUMS 1% sample.

4.4 Smoothing Data of Low Reliability: ACS PUMS Data

Our initial illustration of smoothing with a Rogers-Castro model migration schedule in Fig. 2.4 dealt with data, not from a survey, but from the Swedish national population registration system, which produces age-specific migration rates that more precisely represent the migration age patterns of the population. So far in this chapter, we have dealt with age-specific migration propensities estimated from the

Census 2000 Long-Form Survey, here called the full sample data, and with propensities estimated from a substantially reduced subset of the responses to the Long-Form Survey, the Census 2000 PUMS 1% sample data. At this point, we diverge from the Census 2000 derived estimates and investigate how the same smoothing techniques can be applied to migration age patterns that are calculated with ACS PUMS sample data.

In Section 4.4.1, we discuss how the ACS is a radical departure from the model of the decadal Long-Form Survey, which ended with Census 2000, and we demonstrate how the migration age profiles that result from the ACS are not comparable, substantively, to those from the Census 2000 and from prior decennial censuses. Section 4.4.2 addresses the issue of sampling error in the ACS PUMS samples, and results of simulation experiments are reported showing that the reliability of the estimates of migration propensities is lower in the ACS PUMS sample data than in the Census 2000 PUMS 1% sample data, although the sample sizes are roughly of comparable size. We apply the smoothing techniques to the yearly ACS estimates in Section 4.4.3 using state migration data, and we argue that the smoothing methods are an essential step in reducing irregularities and randomness, regardless of the extent of sampling error in the estimates.

4.4.1 A Comparison of ACS PUMS and Census 2000 Migration Data

In Census 2000 and prior decennial censuses (since 1940), basic migration data were compiled by the U.S. Census Bureau from the Long-Form Surveys, administered as part of the census, and distributed to the public in tables and special cross-tabulated counts. In addition, samples of the data records from individual and household questionnaires were extracted from the Long-Form Survey, and, after special processing to reduce disclosure risks, these data were released to the public as PUMS files. These data files have always been valuable to migration researchers because they contain records at the level of the person and the household, offering maximum analytical flexibility in that migrations can be associated with the characteristics of the persons migrating as well as the characteristics of their households.

Beginning in 2005, the ACS replaced the Long-Form Survey as the primary source of migration data, and the PUMS files that are extracted from the complete annual ACS data have become increasingly important to researchers of U.S. domestic migration data. In contrast to the Long-Form Survey, the ACS is administered in yearly cycles, instead of the 10-year cycles, and each ACS represents roughly 2.5% of the U.S. population, instead of 17% of the population. The complete ACS annual data are still aggregated and distributed to the public in tables and cross-tabulations as was done with the Long-Form Survey data. (The 1-year tabulated ACS products found on the web in the American FactFinder are based on all of the ACS data collected for that year.) However, because of the shortness of the year cycle and because of the low reliability of estimates based on the ACS sample sizes, there are fewer tables published that report annual migration data and the extent to which they are disaggregated by person and household characteristics is significantly diminished.

These changes in U.S. Census Bureau policies, since the implementation of the full ACS, mean that it is not possible to obtain estimates, derived from the complete annual ACS data, which are analogous to those that became available in the special tabulations of the full sample data after the decadal censuses. Here we are referring, specifically, to estimates of age-specific migration flows, marginal flows as well as directional flows, say from one state to another or from one county to another. This policy is justified by U.S. Census Bureau because of the unreliable estimates that inevitably result when the annual ACS data are disaggregated to that level of detail. As an alternative, the U.S. Census Bureau argues that pooling the data from the annual ACS surveys over three and five year periods will improve the reliability of these estimates, and it has begun to release tables of multiyear estimates for these situations (U.S. Census Bureau, 2008).

Some migration researchers counter that the improvements to reliability gained from pooling ACS data over a relatively long period will confound the possibly changing annual patterns (Franklin and Plane, 2006), and for those researchers, who are interested in investigating annual migration patterns in any detail, the annual ACS PUMS sample files are the only source of public data. Ironically, these data suffer even more from reliability problems than the complete ACS data because the PUMS data are extracted to represent 1% of the population instead of 2.5% of the population represented by the complete annual ACS data.

Other aspects of U.S. migration research changed with the full implementation of the ACS in 2005 (U.S. Census Bureau, 2008). For example, the Census 2000 Long-Form Survey migration question asked where the household was living five years ago using a single reference day (April 1, 2000). The question precisely refers to where the person was living on April 1, 1995; therefore, a state out-migrant is someone who was living in a different state in 1995 and lived to be counted by the census takers in 2000. The Census 2000 Long-Form Survey does not detect interstate moves between 1995 and 2000. Only the move between the state of residence on April 1, 1995 and the state of residence on April 1, 2000 is captured in the data.

On the other hand, in the recent years of the ACS, households were asked where they were living one year ago. In the ACS, a state out-migrant, in 2005, for example, is someone who resided in another state in 2004 and lived to be counted by the ACS in 2005. However, because ACS surveys are administered throughout the year, the reference day is imprecise, the household could have moved out of state any time between January, 2004 and December, 2004. The residence rules for the ACS are also quite different than they were for the Census 2000 Long-Form Survey. The Long-Form Survey used the "usual" residence rule which assigns residency based on where the person lives most of the time. The ACS is much more flexible, and assigns residency according to where the person last lived for two months. (See *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know* (U.S. Census Bureau, 2008) for an overview of ACS data.)

As for comparing the migration age patterns that are derived from ACS data to those derived from the Census 2000 Long-Form Survey, some have suggested that the complete ACS yearly data need to be pooled over five years to obtain estimates of similar quality to past census data (Mather, Rivers, & Jacobsen, 2005; Griffin & Waite, 2006). Consider the comparison of age profiles based on interstate rates of migration (in 5-year age groupings) from California to Alaska, Hawaii, Oregon and Washington presented in Fig. 4.6. The Census 2000 migration rates are derived from the Long-Form Survey, i.e., the full sample data, and the ACS data are pooled over the three years: 2005, 2006, 2007. Even when pooled over three years and when the migration propensities are estimated in 5-year age groupings (shown to be more reliable than 1-year age ones in Fig. 4.7) the estimates derived from the pooled ACS PUMS sample files still exhibit more irregularities in comparison with the corresponding Census 2000 data. Moreover, the shapes of the ACS derived patterns are substantially more ragged and more in need of smoothing than the Census 2000 derived propensities.

The age patterns of migration out of California appear to have changed over time, especially from California to Alaska as exhibited in Panel (a). How does one disaggregate the differences in shapes that are due to problems of reliability in the ACS PUMS sample data, to migration patterns changing between the periods (1995–2000) and (2004–2007), to differences in the questions asked by the two surveys, to differences in the time intervals (5 years versus 1 year) or to differences

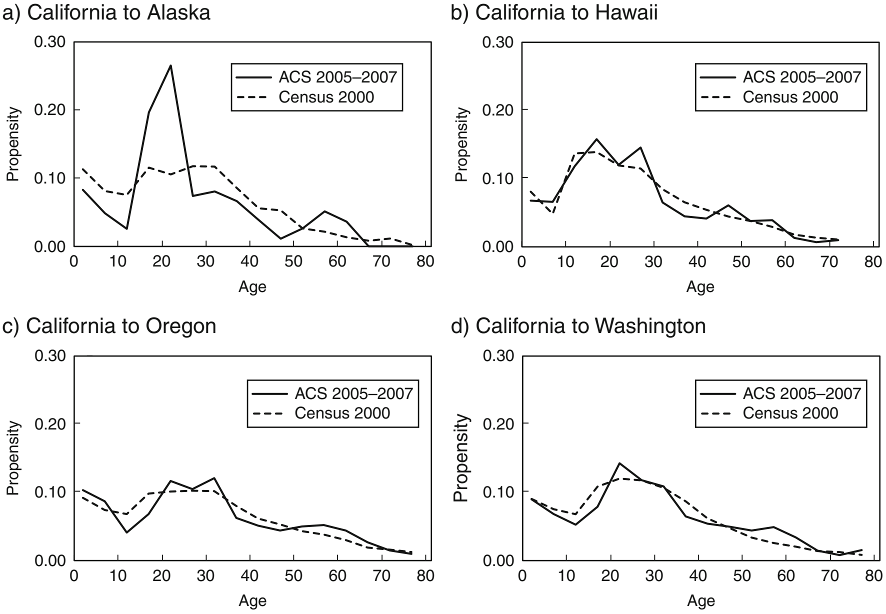


Fig. 4.6 A comparison of interstate migration propensities (5-year age groupings) derived from the ACS 2005–2007 PUMS pooled sample and the Census 2000 full sample

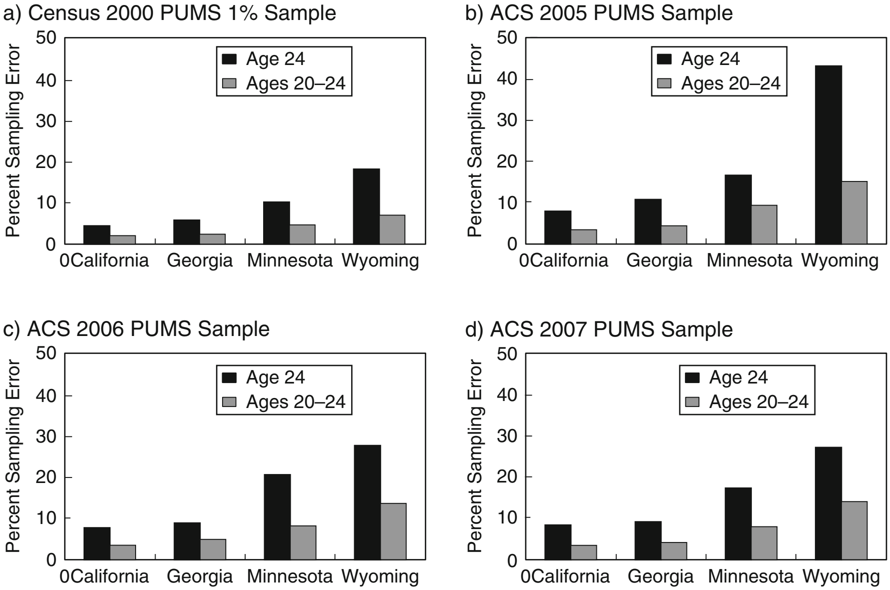


Fig. 4.7 A comparison of the sampling errors of the estimates of age-specific migration propensities derived from the Census 2000 PUMS 1% sample and the ACS 2005–2007 PUMS samples

in the residency rules applied? We conclude that the differences between the two migration data sources are vast, and any conclusions about changes in migration trends based on these comparisons are problematic.

4.4.2 The Reliability of ACS PUMS Estimates

A major difference in the use of ACS data as compared to data from prior decadal censuses is that estimates based on the ACS are expected to be accompanied by some measure of sampling error, whereas, often it has been assumed that estimates derived from census data reflect the population characteristics. Sampling error is the uncertainty associated with an estimate that is based on data gathered from a sample of the population rather than from the full population. One unexpected change in research practices, as a consequence of the changes in direction taken by the U.S. Census Bureau, is that estimates of sampling error are now requested even for the earlier Long-Form Survey as well as the PUMS sample data.

Measures of sampling error give an idea of how precise estimates are, and how appropriate they are for meaningful interpretation. In addition, they are used to tell whether differences over time and space are statistically significant or lie within the bounds of random variation. Measures of sampling error are routinely reported with tables of ACS estimates reported by the U.S. Census Bureau, but for the PUMS files there are no published standard errors. For some estimates derived from the ACS PUMS sample data there is a generalized formula that can be used to calculate standard errors (the fundamental measure of sampling error).

For our purposes, we are interested in the reliabilities of the age-specific conditional survivorships, defined in Section 4.2, which are the basic estimates that undergird a migration age profile. Recall they are denoted either S_i(x) for a 1-year age migration rate from state i, or S_i(x, x+4) for a 5-year age rate, for ages x to x+4. We chose the 1-year age category of 24 and the 5-year age category 20–24 and calculated $S_i(24)$ and $S_i(20, 24)$ for four states, California, Georgia, Minnesota, and Wyoming, representing population sizes decreasing from California (the most populated state) to Wyoming (the least populated state), for the 2005, 2006 and ACS 2007 PUMS samples as well as the Census 2000 PUMS 1% sample.

We used the U.S. Census Bureau's replicate weights method for calculating the standard errors for the ACS PUMS samples (U.S. Census Bureau, 2008, p. 6). If the original weighted sample estimate is denoted $S_i(24)$, then the replicate weights method calls for the calculation to be repeated 80 times from the same sample of 24-year olds, each time weighting each person by one of their provided replicate weights, *rw*1, *rw*2, *rw*3,. . . .,*rw*80, producing 80 new estimates $S_{i1}(24)$, $S_{i2}(24)$, $S_{i3}(24), \ldots, S_{i80}(24)$,. The standard error of the estimate of $S_i(24)$ is expressed in Eq. (4.2) as:

$$SE(\mathrm{S_i}(24)) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(S_{ir} - S_i(24))^2} \tag{4.2}$$

For the Census 2000 PUMS 1% sample the standard errors were estimated by a bootstrap method that drew 80 samples out of 100 weighted observations with replacement.

The coefficient of variation (*CV*) is used to reflect the relative amount of sampling error associated with a sample estimate. The *CV* is calculated as 100 times the ratio of the standard error (*SE*) for an estimate to the estimate itself, or 100* $\left(\frac{SE(S_i(24))}{S_i(24)}\right)$, and here it is referred to as the percent of sampling error.

The results of the *SE* and the *CV* calculations for the PUMS samples of the four surveys are reported in Table 4.3, and they are visually displayed in Fig. 4.7. Dramatic improvements in the reliabilities of migration rates are realized after broadening the age groupings from 1-year age categories to 5-year age groups. The 1-year age estimates have more than twice the percentage of sampling error of the 5-year age estimates. This result is consistent over all four samples and for all four states. A similar result was demonstrated for the Census 2000 PUMS 1% sample data in Section 4.3, and reported in Table 4.1, when the errors in the migration age profiles that resulted from the cublic splined profiles (based on 5-year age propensities) were compared to the errors of the observed migration profiles (based on 1-year age propensities).

Table 4.3 A comparison of the sampling errors of the Census 2000 PUMS 1% and the ACS 2005–2007 PUMS estimates of migration propensities, $S_i(24)$ and $S_i(20,24)$

Census 2000 PUMS 1% Sample	California	Georgia	Minnesota	Wyoming
$S_i(24)$	0.11	0.15	0.12	0.32
SE	0.00	0.01	0.01	0.06
CV	4.42	5.80	10.19	18.16
unweighted *N*	4,880	1,898	636	63
$S_i(20,24)$	0.11	0.16	0.13	0.37
SE	0.00	0.00	0.01	0.03
CV	2.03	2.39	4.67	7.00
unweighted *N*	22,187	8,399	2,633	334
ACS 2005 PUMS Sample				
$S_i(24)$	0.04	0.05	0.07	0.09
SE	0.00	0.01	0.01	0.04
CV	7.97	10.79	16.68	43.14
unweighted N	4,073	1,720	539	54
$S_i(20,24)$	0.04	0.05	0.06	0.14
SE	0.00	0.00	0.01	0.02
CV	3.49	4.44	9.42	15.17
unweighted *N*	19,823	8,556	2,562	291
ACS 2006 PUMS Sample				
$S_i(24)$	0.04	0.06	0.05	0.13
SE	0.00	0.00	0.01	0.04
CV	7.73	8.87	20.58	27.63
unweighted *N*	4,212	1,720	526	62
$S_i(20,24)$	0.04	0.05	0.05	0.13
SE	0.00	0.00	0.00	0.02
CV	3.50	4.91	8.19	13.63
unweighted *N*	21,135	9,415	2,584	320
ACS 2007 PUMS Sample				
$S_i(24)$	0.03	0.06	0.06	0.15
SE	0.00	0.01	0.01	0.04
CV	8.40	9.17	17.33	27.17
unweighted *N*	4,448	2,020	530	74
$S_i(20,24)$	0.03	0.06	0.07	0.15
SE	0.00	0.00	0.01	0.02
CV	3.45	4.16	7.98	14.07
unweighted *N*	21,511	9,785	2,554	324

Another result that is consistent for all four samples, and reported in Fig. 4.7, is that the percentage of sampling error increases as the population size decreases. The four states selected are decreasing in population size from left to right in Fig. 4.7. In 2005, California was the most populated state (population 35,340,566), Georgia was the eleventh most populated (population 8,811,648), Minnesota ranked 24 in population (4,969,152), and Wyoming was the least populated state (494,170).

The final result displayed in Fig. 4.7, and perhaps the most significant, is that the sampling errors of the estimates from ACS PUMS sample data are larger than the sampling errors from the Census 2000 PUMS 1% sample data. This is true for the 1-year and 5-year age propensities and for all ACS yearly samples. However, the percentage of sampling error was largest in the ACS 2005 PUMS sample data, which was the first year of the full implementation of the ACS. It appears that after 2005, the sampling error leveled off because, as displayed in Fig. 4.7, the sampling error results are very similar for 2006 and 2007 for all of the estimates examined. It is also important to note that, on average, the percentage differences in sampling errors between the Census 2000 PUMS 1% sample and the ACS PUMS samples, for the estimates of the 5-year age propensities, increased linearly as the sample size decreased. For California, the average percentage increase in error between the Census 2000 PUMS 1% sample and the ACS PUMS samples was 72%, for Georgia (89%), for Minnesota (83%) and for Wyoming (104%).

The U.S. Census Bureau offered guidelines for interpreting coefficients of variation associated with ACS PUMS sample estimates (Robinson, 2009). These guidelines suggest that estimates with coefficients of variation that are less than 15% can be considered reliable, and unreliable if the coefficient of variation is greater than 30%. Only one of the coefficients of variation reported in Table 4.3, for the aged 20–24 migration propensity estimates, is greater than 15%, and that is the one for Wyoming in 2005.

4.4.3 Results for the State Data

Given the substantive differences between the Census 2000 Long-Form Survey and the ACS in design and in the migration questions asked, as well as the apparent decrease in reliability associated with the ACS PUMS sample data, it is reasonable to question the effectiveness of our proposed smoothing procedures when applied to these data. In this section of the chapter, we demonstrate the smoothing techniques, focusing on the ACS 2005 PUMS data, which has the least reliable estimates of migration age propensities when compared to the other years of ACS PUMS data collected since then (see Fig. 4.7 and Table 4.3).

In Figs. 4.5 and 4.7 that the more populated states have larger PUMS sample sizes, and, consequently, their estimates are more reliable. Therefore, we begin with the ACS 2005 PUMS estimates of migration age propensities for California, the most populated state in 2005 (population = 35,340,566). Figure 4.8 visually displays the sequence of the smoothing steps, beginning with the somewhat jagged profile of migration propensities for 1-year age groups in Panel (a). Panel (b) shows the histogram of migration age propensities of the 5-year age groupings, which roughly outlines a typical migration profile with only one irregularity. Just to the right of the career migration peak, the estimated propensities for the age groups 25–29 and 30–34 do not gradually lower, but instead level off. This is even more apparent in Panel (c) where the results of the cubic splined interpolation are displayed. Fortunately, the model migration schedule that was fitted to the splined profile, displayed in Panel (d), eliminated the irregularity and produced a profile that conforms to a shape that is a more acceptable representation of a population migration schedule.

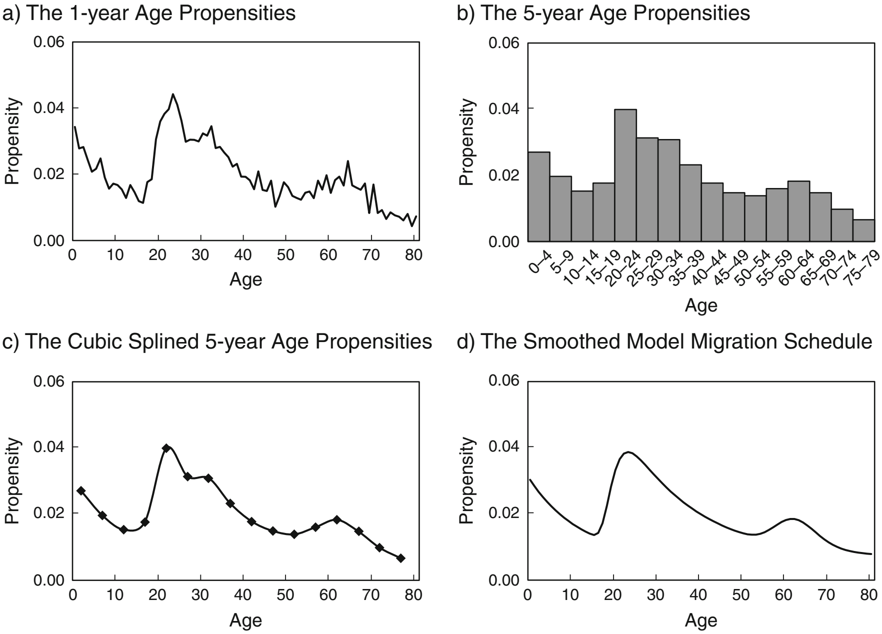


Fig. 4.8 A demonstration of the smoothing procedures for the California migration age propensities derived from the ACS 2005 PUMS sample

In contrast to the California example demonstrated in Fig. 4.8, the smoothing procedures for Wyoming are illustrated in the panels of Fig. 4.9. Wyoming was the least populated state in 2005 (population = 494,170). The estimated 1-year migration age propensities for Wyoming are reported in Panel (a). They oscillate wildly and there are discontinuities where there are data missing for certain 1-year age categories. The migration propensities for the 5-year age groupings, presented in Panel (b), produce a profile that is not recognizable as a migration age pattern. The same is true of the cubic splined profile exhibited in Panel (c), particularly because of the lack of an infant migration peak in the 0–4 age category. Closer examination of the propensities of the 1-year age categories in Panel (a) shows there is a spike in the estimated migration propensity for the age 0 category, but the combined 5-year propensity for ages 0–4 is atypically lower than the propensity for the 5–9 age grouping. In addition, substantial oscillation remains in the age groupings older than that of the labor peak (ages 20–24). The final result of the smoothing procedures, displayed in Panel (d) is a simple, but regular, model schedule that imposes the “best fitting” standard shape on the irregular data.

These two demonstrations clearly show the problems of the irregularities in the migration propensities derived from the ACS PUMS data, for the least populated state (Wyoming) as well as the most populated state (California), and they show the effectiveness of the smoothing procedures. Because there is no “gold standard” to represent the “true” migration age schedules for the years of the ACS surveys, like those derived from the Long-Form Survey in Census 2000, it is difficult to

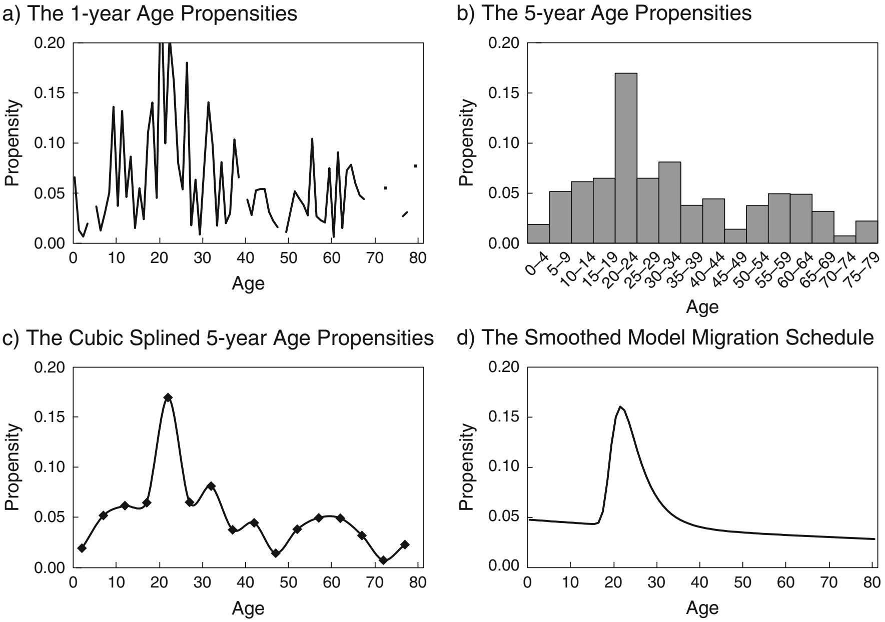


Fig. 4.9 A demonstration of the smoothing procedures for the Wyoming migration age propensities derived from the ACS 2005 PUMS sample

gauge the improvements to the reliability of the estimated migration schedules gained by the smoothing procedures. For the experiments done with Census 2000 PUMS 1% sample data we were able to compare the different stages of the smoothing process with the model schedules derived from the Census 2000 full sample data, and these comparisons were summarized with the *MAPE* statistics reported throughout Section 4.3. Those *MAPE* (Mean Absolute Percent Error) statistics can be interpreted as measures of reliability (with larger *MAPE* values indicating lower reliability) because they capture the degree of disparity between the profiles derived from the sample estimates and the "true" population profiles.

For the experiments with ACS PUMS sample data we use the Mean Absolute Percent Difference (*MAPD*) index, which is calculated like the *MAPE*, and here is used to assess the difference between the splined profiles (based on the cubic spline interpolation of the 5-year propensities) and the Rogers-Castro model migration schedules that result from the final step of the smoothing procedures. The *MAPD* cannot be interpreted as a measure of reliability, or average percent error, because it is used simply to gauge the irregularities of the splined profile with respect to the smoothed model schedule. It is important to note that both of the schedules involved in the *MAPD* calculation (i.e., the spline and the model schedule) are derived from the ACS PUMS sample data and neither represents a "true" population migration schedule.

For example, the *MAPD* that resulted from the California application of the smoothing procedures, displayed in Fig. 4.8, is 4.06 (summarizing the contrast between the splined profile in Panel (c) and the model schedule in Panel (d)). In other words, on average, over all ages, the splined profile exhibits a 4.06% departure from the regularity of the model schedule. The comparable *MAPD* resulting from the Wyoming experiment, reported in Fig. 4.9, is 31.62, which suggests that, on average, over all the ages, there is a 31.62% difference between splined profile and the model schedule.

The *MAPD*s measuring the contrast between the splined profiles and the model schedules derived from the ACS 2005–2007 PUMS data are reported for all states in Table 4.4. They can be used to assess the quality of the estimates of the migration propensities for the 5-year age groupings. For example, a smaller *MAPD* value indicates that the splined profile, interpolated from the 5-year age propensities, is quite similar to the smooth model schedule. Conversely, a larger *MAPD* value suggests that the splined profile is very irregular and departs severely from the smoothed and regular fitted model schedule.

State *MAPD* values are negatively associated with state population size (and sample size). All three statistics are reported in Table 4.4 for the 50 states (and the District of Columbia), for all three years of the ACS. The average *MAPD*s are reported in Fig. 4.10 for states within population size categories, representing the results of the smoothing procedures on the Census 2000 PUMS 1% data, in Panel (a), and on the ACS 2005 PUMS data, in Panel (b). In both samples, the average *MAPD*s increase as state population size decreases. Clearly, the average *MAPD*s are larger for the ACS 2005 PUMS sample data than for the Census 2000 PUMS 1% sample data, and this disparity is most dramatic for the very least populated states, for example, those with less than 1 million people.

It seems plausible that the *MAPD*, which gauges irregularity in the sample estimates manifested in the splined profiles, is related to the *MAPE*, which measures error (unreliability) in the smoothed migration schedule as compared to the true migration schedule. However, the *MAPE* is not attainable from the ACS smoothing experiments, and it *is* available from the Census 2000 PUMS 1% sample experiments. These average *MAPE*s are reported in the right most column of Table 4.2, and they are displayed again in Panel (a) of Fig. 4.10. The results reported in Panel (a) reveal that the model schedules resulting from the smoothing procedures are closer to the "true" model schedules than the splined profiles are to the smoothed model schedules. In addition, the reliability of the smoothed model schedules shows the same familiar negative association with population size.

To gain some understanding of the reliabilities of the smoothed model schedules that resulted from the experiments with the ACS 2005 PUMS sample data, we borrowed the fraction of the *MAPE* to the *MAPD* from the Census 2000 PUMS 1% sample experiments on each state, and we multiplied them by the *MAPD* that resulted from the ACS 2005 PUMS data experiments. The averages of these values are reported in Table 4.5 and labeled "Average Estimated *MAPE*." This finding is based on the assumption that there is a consistent relationship between the regularity in the sample based splined profiles and the reliability of the model schedules that result from the smoothing procedures. By invoking this assumption, we gain an assessment of the reliability of the model schedules resulting from the smoothing

Table 4.4 State population size, unweighted sample size, and the *MAPD*s derived from the ACS 2005, 2006 and 2007 PUMS sample data

	Population			Sample size			*MAPD*		
State	2005	2006	2007	2005	2006	2007	2005	2006	2007
Alabama	4,448,075	4,599,030	4,627,851	45,534	47,018	47,329	14.68	12.59	16.62
Alaska	658,002	670,053	683,478	6,129	6,327	6,386	24.04	31.77	18.46
Arizona	5,806,266	6,166,318	6,338,755	58,315	60,195	61,436	8.97	8.47	10.32
Arkansas	2,694,665	2,810,872	2,834,797	27,399	28,343	28,627	17.42	14.21	13.77
California	35,340,566	36,457,549	36,553,215	334,885	345,723	347,958	4.06	7.09	3.68
Colorado	4,540,639	4,753,377	4,861,515	46,094	48,020	48,256	12.57	8.06	6.21
Connecticut	3,365,768	3,504,809	3,502,309	33,867	35,070	35,198	17.30	11.22	15.91
Delaware	825,598	853,476	864,764	7,933	8,115	8,370	29.88	31.76	10.93
District of Columbia	508,572	581,530	588,292	5,187	5,577	5,476	27.39	11.98	24.98
Florida	17,363,653	18,089,889	18,251,243	177,000	185,309	185,538	5.87	6.12	5.39
Georgia	8,811,648	9,363,941	9,544,750	87,534	91,896	93,300	12.26	8.75	10.08
Hawaii	1,258,528	1,285,498	1,283,388	12,743	12,891	13,102	22.06	12.89	13.06
Idaho	1,408,650	1,466,465	1,499,402	14,353	14,931	15,165	18.80	18.79	19.26
Illinois	12,441,864	12,831,970	12,852,548	123,074	126,613	127,458	9.17	5.19	8.18
Indiana	6,081,212	6,313,520	6,345,289	63,278	65,054	65,776	14.99	9.26	16.76
Iowa	2,848,266	2,982,085	2,988,046	29,629	30,883	30,951	12.71	18.32	11.94
Kansas	2,669,699	2,764,075	2,775,997	27,462	28,168	28,347	21.56	17.05	23.09
Kentucky	4,065,635	4,206,074	4,241,474	41,498	42,429	42,962	10.35	12.62	13.87
Louisiana	4,387,181	4,287,768	4,293,204	43,956	40,901	42,073	5.26	5.13	10.61
Maine	1,282,474	1,321,574	1,317,207	12,440	12,649	12,631	20.33	18.72	15.96
Maryland	5,453,441	5,615,727	5,618,344	54,290	55,683	56,247	11.73	20.78	12.73
Massachusetts	6,200,944	6,437,193	6,449,755	62,695	64,673	64,330	12.23	10.93	9.90
Michigan	9,857,477	10,095,643	10,071,822	99,784	101,355	101,093	11.21	11.54	10.17
Minnesota	4,969,152	5,167,101	5,197,621	50,857	52,219	52,854	6.47	6.88	15.12
Mississippi	2,830,388	2,910,540	2,918,785	28,354	28,945	29,066	9.00	10.74	17.42
Missouri	5,632,603	5,842,713	5,878,415	57,884	59,696	60,095	10.29	7.03	10.52
Montana	897,367	944,632	957,861	8,715	9,052	9,070	23.57	17.91	17.47

Table 4.4 (continued)

	Population			Sample size			*MAPD*		
State	2005	2006	2007	2005	2006	2007	2005	2006	2007
Nebraska	1,706,343	1,768,331	1,774,571	17,442	18,063	17,960	17.79	15.49	12.19
Nevada	2,376,017	2,495,529	2,565,382	23,538	24,858	25,688	12.31	6.83	10.06
New Hampshire	1,271,897	1,314,895	1,315,828	12,758	12,818	13,183	17.95	14.05	21.52
New Jersey	8,524,868	8,724,560	8,685,920	83,991	86,190	86,527	6.54	13.99	11.57
New Mexico	1,886,789	1,954,599	1,969,915	18,272	18,637	18,593	17.28	13.17	17.05
New York	18,679,211	19,306,183	19,297,729	181,406	187,143	187,057	6.36	7.36	4.33
North Carolina	8,397,785	8,856,505	9,061,032	85,611	89,124	91,320	11.53	7.55	9.73
North Dakota	621,063	635,867	639,715	6,505	6,699	6,689	31.10	20.76	23.73
Ohio	11,146,050	11,478,006	11,466,917	114,707	117,593	117,543	6.38	4.17	11.54
Oklahoma	3,429,974	3,579,212	3,617,316	34,683	35,781	36,366	13.36	12.62	13.18
Oregon	3,560,922	3,700,758	3,747,455	35,485	36,499	37,076	10.53	10.22	8.35
Pennsylvania	11,948,862	12,440,621	12,432,792	121,424	124,455	125,637	10.20	9.45	7.84
Rhode Island	1,033,284	1,067,610	1,057,832	10,184	10,576	10,488	18.80	25.31	25.34
South Carolina	4,127,391	4,321,249	4,407,709	41,956	43,829	44,409	10.62	11.29	20.18
South Dakota	755,152	781,919	796,214	7,667	8,044	7,968	28.09	31.37	40.97
Tennessee	5,816,359	6,038,803	6,156,719	59,376	61,139	61,704	9.75	15.71	10.24
Texas	22,250,152	23,507,783	23,904,380	217,617	226,724	230,817	10.51	5.88	5.46
Utah	2,452,149	2,550,063	2,645,330	24,749	25,746	26,311	18.39	17.31	18.98
Vermont	609,857	623,908	621,254	5,896	6,183	6,164	38.72	30.74	20.49
Virginia	7,320,848	7,642,884	7,712,091	73,509	76,649	77,525	12.57	12.21	9.24
Washington	6,157,786	6,395,798	6,468,424	61,520	63,524	64,414	11.11	12.29	9.57
West Virginia	1,781,817	1,818,470	1,812,035	17,771	18,446	18,390	13.78	13.13	19.87
Wisconsin	5,401,740	5,556,506	5,601,640	56,368	57,987	58,435	10.04	13.28	17.72
Wyoming	494,170	515,004	522,830	5,056	5,299	5,304	31.62	33.86	26.04

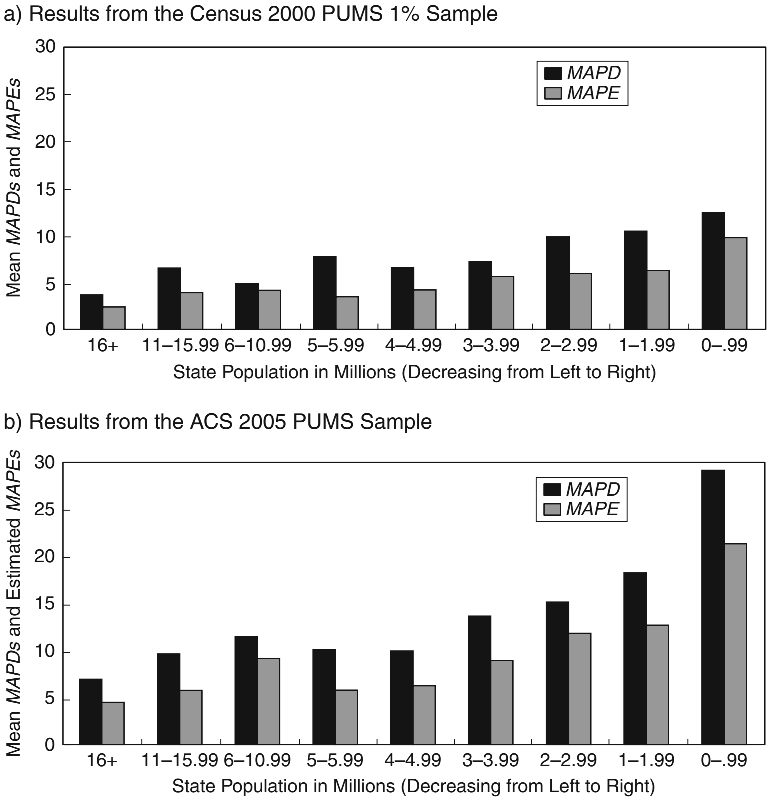


Fig. 4.10 A comparison of the regularity (*MAPD*) and the reliability (*MAPE*) results from the smoothing procedures performed on the Census 2000 PUMS 1% sample and the ACS 2005 PUMS sample by state population size

procedures performed on the ACS 2005 PUMS data. In addition, if we apply the guideline that sampling errors less than 15% suggest reliable ACS PUMS sample estimates, which was offered by the U.S. Census Bureau (Robinson, 2009), we can conclude that the smoothing procedures provide reliable model migration schedules for all states, except the very least populated states with less than 1 million persons, which have an average estimated *MAPE* value of 21.46.

4.5 Log-Linear Smoothing of Spatial and Age Patterns in Migration Flow Tables

In this section, we show how the unsaturated log-linear model, introduced in Chapter 3, can be used to smooth the spatial and age structures in migration flow tables. The model migration schedule approach described in the previous sections

Table 4.5 A comparison of the regularity (*MAPD*) and the reliability (*MAPE*) results from the smoothing procedures performed on the Census 2000 PUMS 1% sample and the ACS 2005 PUMS sample by categories of state population size (decreasing)

	Census 2000 PUMS 1% Sample			ACS 2005 PUMS Sample		
State population size categories (in millions)	N	Average *MAPD*	Average *MAPE*	N	Average *MAPD*	Average Estimated *MAPE*
16+	3	3.65	2.38	3	6.98	4.51
11–15.99	3	6.51	3.90	2	9.68	5.79
6–10.99	5	4.85	4.15	8	11.55	9.22
5–5.99	6	7.74	3.47	5	10.16	5.81
4–4.99	5	6.57	4.19	6	9.99	6.29
3–3.99	6	7.19	5.61	3	13.73	9.00
2–2.99	5	9.82	5.94	6	15.23	11.89
1–1.99	9	10.43	6.27	8	18.35	12.76
0–0.99	8	12.41	9.74	8	29.30	21.46

can be considered as a “bottom-up” approach that smoothes the age profile of each flow in a migration flow table. The log-linear model, on the other hand, can be viewed as a “top-down” approach in which higher-order marginal totals of, for example, an origin-by-destination-by-age table of migration flows are assumed to be more reliable (and regular) than lower-order marginal totals or cell values. Here, the data may be smoothed by removing, for example, the two-way and three-way interaction terms from the saturated model. In this section, we first focus on log-linear models for smoothing spatial patterns of migration (Section 4.5.1), followed by log-linear models for smoothing age patterns of origin-destination-specific flows (Section 4.5.2). In Section 4.5.3, we show how model migration schedules may be combined with log-linear models to form hybrid models that may lead to further improvements in terms of both fit and parsimony.

4.5.1 Spatial Patterns of Migration between the Nine U.S. Divisions

We explore the use of log-linear methods for smoothing the spatial patterns of migration by focusing on two applications. The first represents a situation where several periods of migration data are available, and the aim is to smooth trends in the observed patterns over time. The second represents a situation where only a single period of migration data is available. Here, the aim is to remove some of the irregularities caused by small sample size or to remove some of the effects of specific period abnormalities in the data. We use U.S.-born migration between the nine divisions in the U.S. for an illustration. These data were obtained from the 1980, 1990 and 2000 censuses and are set out in Table 4.6.

Table 4.6 U.S.-born migration flows (in thousands) between the nine U.S. Census divisions, 1975–1980, 1985–1990, and 1995–2000

	Destination									
Origin	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific	Total
(a) 1975–1980										
New England	0	182	79	29	270	26	60	65	130	841
Middle Atlantic	291	0	265	74	1,052	81	179	183	319	2,444
East North Central	82	185	0	311	719	290	329	321	435	2,671
West North Central	25	43	247	0	151	56	260	240	220	1,242
South Atlantic	132	356	358	118	0	364	337	159	309	2,134
East South Central	15	40	201	53	312	0	177	42	85	925
West South Central	26	52	126	141	203	136	0	175	239	1,098
Mountain	24	46	104	139	101	32	204	0	449	1,098
Pacific	65	113	189	179	255	77	319	567	0	1,764
Total	660	1,017	1,570	1,045	3,062	1,060	1,865	1,752	2,186	14,218
(b) 1985–1990										
New England	0	178	65	22	318	19	31	44	100	777
Middle Atlantic	272	0	194	50	1,079	61	95	104	222	2,076
East North Central	82	177	0	271	774	249	209	218	313	2,294
West North Central	32	53	258	0	192	54	193	215	202	1,199
South Atlantic	154	380	380	104	0	326	220	134	290	1,987
East South Central	19	40	175	46	407	0	134	37	73	931
West South Central	58	109	249	227	480	216	0	234	354	1,927
Mountain	41	70	153	153	173	42	183	0	575	1,388
Pacific	88	141	212	140	349	78	229	524	0	1,762
Total	746	1,148	1,686	1,012	3,771	1,047	1,294	1,509	2,128	14,341

Table 4.6 (continued)

	Destination									
Origin	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific	Total
(c) 1995–2000										
New England	0	167	61	22	298	23	41	59	100	771
Middle Atlantic	245	0	199	54	1,084	74	105	145	191	2,097
East North Central	68	161	0	297	674	280	223	273	241	2,217
West North Central	25	48	270	0	185	63	205	215	145	1,157
South Atlantic	168	437	413	139	0	393	314	215	301	2,380
East South Central	18	40	185	47	379	0	159	54	67	947
West South Central	37	76	184	188	358	179	0	235	226	1,482
Mountain	43	72	154	166	197	53	222	0	472	1,379
Pacific	92	151	230	180	397	101	310	766	0	2,227
Total	696	1,150	1,696	1,093	3,573	1,165	1,581	1,962	1,741	14,657

A saturated log-linear model for analyzing migration flow tables over time, such as those set out in Table 4.6, is specified as:

$$\ln\left(\hat{n}_{ijt}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_t^T + \lambda_{ij}^{OD} + \lambda_{it}^{OT} + \lambda_{jt}^{DT} + \lambda_{ijt}^{ODT}, \quad i \neq j \tag{4.3}$$

where O, D and T denote origin, destination and time, respectively. For smoothing these data over time, we consider two unsaturated log-linear models as candidates. The first model includes all the main effects and a two-way interaction term between origin and destination, i.e.,

$$\ln\left(\hat{n}_{ijt}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_t^T + \lambda_{ij}^{OD}, \quad i \neq j. \tag{4.4}$$

This model adjusts the average (or pooled) origin-destination-specific patterns of migration up or down according to the total level of migration observed for a particular period. The second model includes all two-way interactions and is specified as,

$$\ln\left(\hat{n}_{ijt}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_t^T + \lambda_{ij}^{OD} + \lambda_{it}^{OT} + \lambda_{jt}^{DT}, \quad i \neq j. \tag{4.5}$$

This model is more complicated in that it allows the proportional distributions of the origin and destination marginal terms to vary over time. The only term not included is the three-way interaction between origin, destination, and time. The likelihood ratio statistic for the first model, Eq. (4.4), was 903.2, with 142 residual degrees of freedom, whereas for the second model, Eq. (4.5), it was 55.1 with 110 residual degrees of freedom. The second model clearly performed better and, in fact, nearly matched the observed data perfectly, implying that it only very slightly smoothed the spatial patterns over time. Thus, in our analysis below, we focus only on the simpler model.

To illustrate the smoothing of spatial patterns resulting from the relatively simple model specified in Eq. (4.4), consider the flows from the New England and Pacific Divisions set out in Fig. 4.11. Here, we find that the smoothed versions of the migration flows from the New England and Pacific Divisions do not contain any of the irregular patterns exhibited, for example, in the New England to South Atlantic or Pacific to West South Central flows. The model essentially removes the "bumpiness" in the data resulting in patterns that follow smooth trends over time with not much change in the levels. Notice, however, that the smoothed data all exhibit increases in the origin-destination movements over time. This is a consequence of the model specified in Eq. (4.4), which adjusts a single (pooled) set of origin-destination flows according to the overall level of migration observed in each period. For these data, the overall levels increased from 14.2 million during the 1975–1980 period to 14.3 million in the 1985–1990 period to 14.7 million in the 1995–2000 period. A more realistic model would include the additional two-way interactions between origin and time and destination and time, such as those specified in Eq. (4.5).

What about the data situation where only a single period of migration flows are available? How might one smooth these data if there are believed to be some irregularities or abnormalities in the observed patterns? In this case, the migration data

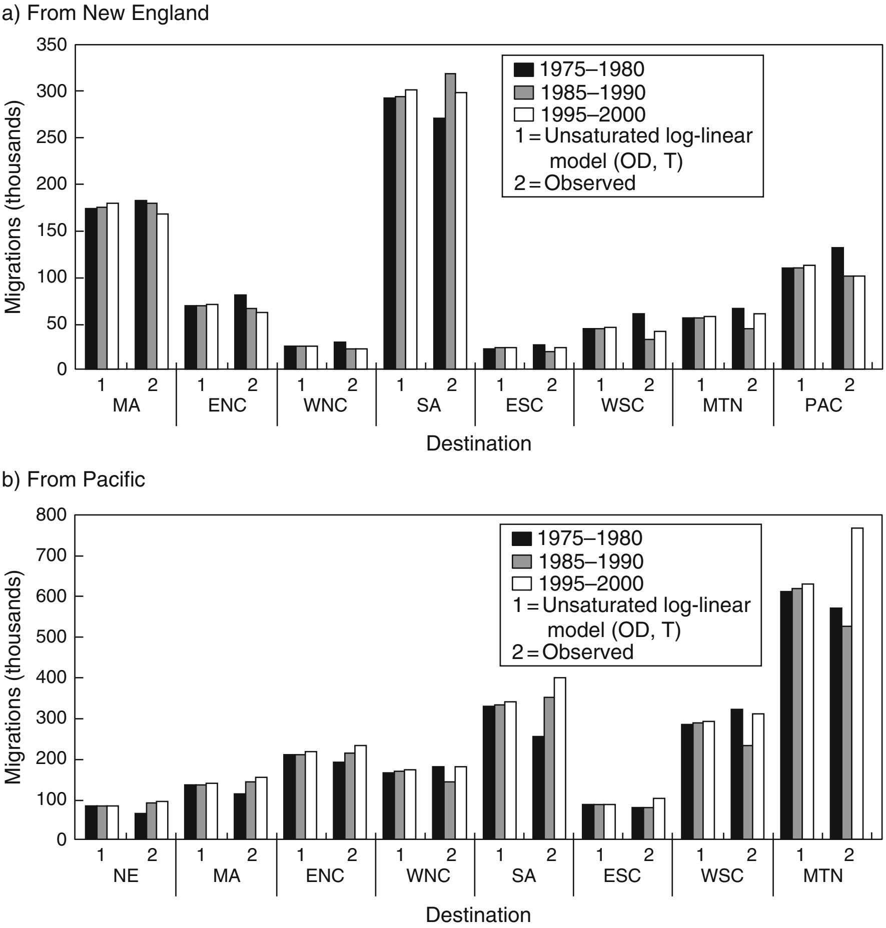


Fig. 4.11 Smoothed and observed flows of migration (in thousands) from New England and Pacific to the other divisions in the U.S.: 1975–1980, 1985–1990, and 1995–2000

can be smoothed by relating the observed flows to some theoretical (or expected) set of flows, e.g., the simplest being a set of flows produced under the assumption of "quasi-independence." We illustrate this for the 1995–2000 migration data set out in Panel (c) Table 4.6. First, the expected flows are predicted under the assumption of quasi-independence between origin and destination, i.e., using the following unsaturated log-linear model with structural zeros:

$$\ln\left(\hat{n}_{ij}\right) = \lambda + \lambda_i^O + \lambda_j^D, \quad i \neq j. \tag{4.6}$$

Second, the average of these predicted flows and the observed set of flows are used to obtain the smoothed set of flows. The results from this exercise are set out for flows from the New England and Pacific Divisions in Fig. 4.12. Here, we see

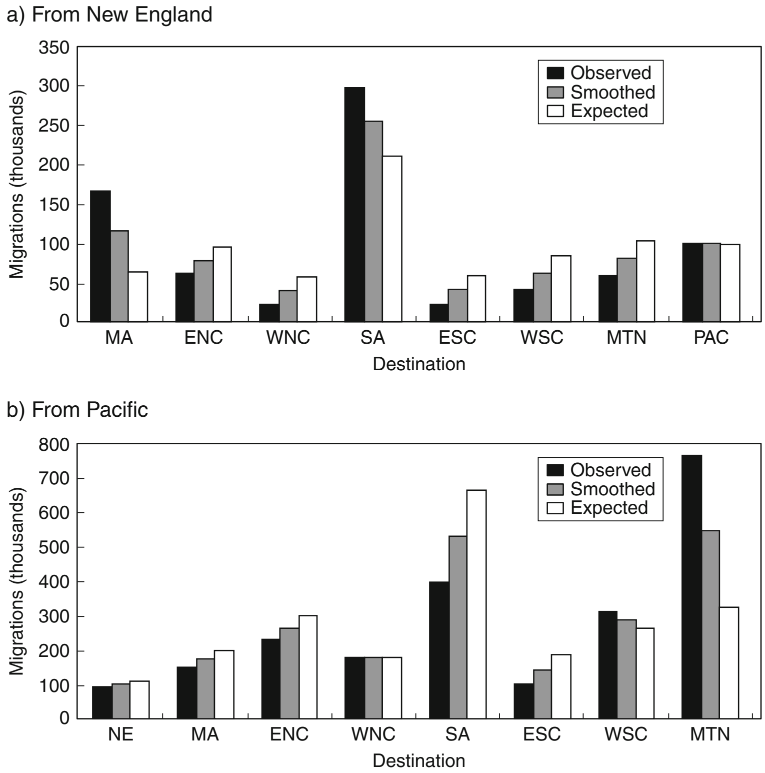


Fig. 4.12 Observed, smoothed, and expected flows of migration (in thousands) from New England and Pacific to the other divisions in the U.S., 1995–2000

that the smoothed set of flows are reduced when the observed patterns are higher than the expected patterns (e.g., New England to South Atlantic) and are increased when the observed patterns are lower (e.g., Pacific to South Atlantic).

The assumption of quasi-independence results in large predicted flows between areas that send or receive large numbers of migrants and, likewise, small predicted flows between areas that send or receive small numbers of migrants. This assumption is very simplistic and ignores other important migration factors, such as distance, contiguity or the relative incomes in each area. An improved smoothing model could bring in these factors (and others) to derive an expected table of migration flows from which the observed data could then be smoothed.

4.5.2 Age Patterns of 1995–2000 Migration from Colorado to Other U.S. States and Divisions

The above log-linear models for smoothing origin-by-destination tables may be extended to include age and other categorical variables of interest. In this chapter, we have shown that irregularities may be caused by period-specific variation or by sample size problems, such as those exhibited in the interstate migration data from the ACS PUMS samples. In this section, we focus on smoothing the irregularities caused by small sample size. To illustrate, we smooth the age-specific migration flows from Colorado to twelve states in the West region and seven divisions outside the West region during the 1995–2000 period. These data were obtained from the Census 2000 PUMS 5% sample data. Since we also have the full sample of Census 2000 migration data, we can test whether smoothing the age patterns of migration actually improves the accuracy of the data. To tie into earlier sections in this chapter, we can also compare the accuracy of the log-linear approach with the 7-parameter model schedule approach (there were no retirement peaks or upward slopes exhibited in these flows).

In order to compare the log-linear and model schedule approaches for smoothing age patterns of migration, we assume that the *aggregate* origin-destination-specific flows, n_{ij+}, are reliable. This assumption allows us to fit model schedules to the age compositions of migration (i.e., $N_{ij}(x) = n_{ijx}/n_{ij+}$) that can then be rescaled to match the aggregate origin-destination totals. In the log-linear model, we include the two-way interaction term between origin and destination. Thus, we can compare the fits resulting from both the log-linear model and the model schedules.

The unsaturated log-linear model used to smooth the age-specific migration data is specified as:

$$\ln\left(\hat{n}_{ijx}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA}, \quad i \neq j. \tag{4.7}$$

This model includes all two-way interactions and thus relies on the marginal age structures (OA and DA) to obtained smoothed estimates of the age-specific migration flows. As for the model migration schedule approach, the smoothed counts of age-specific migration, $\hat{n}_{ijx}$, were obtained by multiplying $N_{ij}(x)$ by n_{ij+}, where $N_{ij}(x)$ and $\hat{n}_{ijx}$ denote the estimated (smoothed) age-specific compositions and counts, respectively. This allowed us to maintain the aggregate levels of origin-destination-specific migration that were considered to be reliable and consistent with the log-linear model specified in Eq. (4.7).

The goodness-of-fits (R^2) comparing the full sample Census 2000 data with the unsaturated log-linear model estimates, the model migration schedules and the actual PUMS 5% sample data are set out in Table 4.7. Here, we find that, on average, the unsaturated log-linear model, Eq. (4.7), produced the best results, both in terms of individual fits and in terms of variance. However, there were many instances where the Rogers-Castro model migration schedules performed better, and even a few instances where the unaltered PUMS 5% sample data represented the best correspondence to the full sample data.

For a better understanding of how the PUMS 5% sample data, model schedule fits, and unsaturated log-linear model fits differ, a selection of flows representing age-specific migration from Colorado to Arizona, Wyoming, and Hawaii are set out in Fig. 4.13, with each flow representing a different situation of best fit. The PUMS 5% sample data best represented the Colorado to Arizona flow because it both corresponded to the full sample pattern and captured the unusual dip after the labor force peak. The model schedule simply fitted a line through the dip. And the

Table 4.7 Goodness-of-fit statistics (R^2) comparing the age compositions of Colorado out-migrants from the full sample Census 2000 data with the unsaturated log-linear model estimates, the Rogers-Castro model migration schedules, and the PUMS 5% sample data

Destination	Log-linear model	Model schedule	PUMS 5% sample
Alaska	**0.98**	0.89	0.94
Arizona	0.94	0.88	**0.97**
California	**1.00**	0.98	0.99
Hawaii	**0.94**	0.88	0.86
Idaho	**0.96**	0.96	0.86
Montana	**0.95**	0.81	0.77
Nevada	0.91	**0.97**	0.94
New Mexico	**0.94**	0.94	0.94
Oregon	**0.98**	0.97	0.97
Utah	**0.98**	0.95	0.98
Washington	0.98	0.98	**0.99**
Wyoming	0.92	**0.95**	0.83
New England	0.98	**0.99**	0.98
Middle Atlantic	0.98	**0.99**	0.99
East North Central	**1.00**	0.98	0.98
West North Central	0.98	**0.99**	0.98
South Atlantic	0.97	**1.00**	0.99
East South Central	0.97	**0.98**	0.98
West South Central	0.99	0.98	**0.99**
Mean	**0.96**	0.95	0.94
Min	0.91	0.81	0.77
Max	1.00	1.00	0.99
SD	0.03	0.05	0.07

Note: Best fits are set in boldface

log-linear model produced a sharper labor force peak, which came from the marginal age structures of out-migration from Colorado and in-migration to Arizona. For the Colorado to Wyoming flow, the PUMS 5% sample data exhibited irregular patterns compared to the full sample data. Here, the model migration schedule was able to produce a curve closer to the full sample flow by fitting a curved line through the irregular patterns. This was, however, not the case for the Colorado to Hawaii flow, where the model schedule fit was unable to represent the sharp labor force peak because it was not captured in the PUMS 5% sample data. In this case, the log-linear model, again relying on marginal age structures with a sharp labor force peak, performed better.

4.5.3 Age Patterns of ACS 2004 Migration between States in the U.S. West Region

The previous section demonstrated the power of the unsaturated log-linear model to smooth age patterns in migration flow tables. This model assumes that the

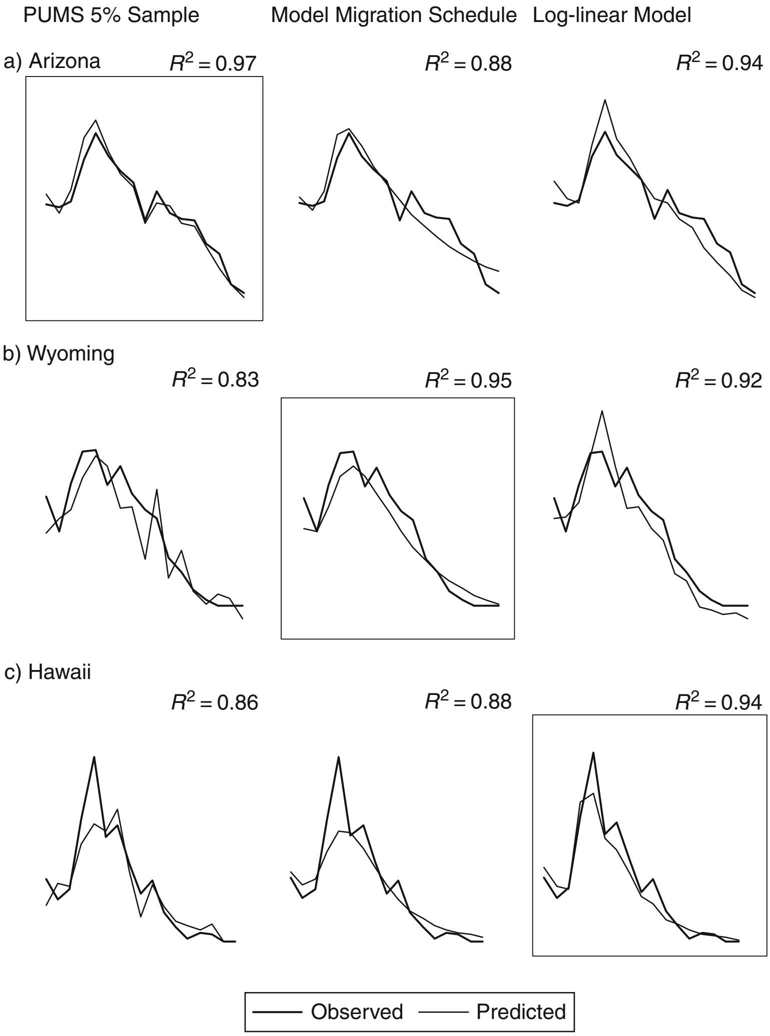


Fig. 4.13 Census 2000 observed (full sample) versus predicted age-specific migration flows from Colorado to Arizona, Wyoming, and Hawaii using the PUMS 5% sample, the Rogers-Castro model migration schedule fits to the PUMS 5% sample, and the unsaturated log-linear model fits to the PUMS 5% sample

marginal age structures in these tables are reliable. But what if they are not? And what if the data are so poor that model schedules cannot be used to smooth the origin-destination-specific flows either? This section shows how the marginal age structures in a migration flow table can be smoothed and used in an offset to obtain a complete set of smoothed migration flows between all regions of interest. For illustration, we use age-specific migration between states in the U.S. Pacific Division (i.e., Alaska, California, Hawaii, Oregon and Washington), obtained from the ACS

2004 PUMS sample, representing a very poor data situation and, to be fair to the ACS, one that would not likely be used for analyses of interstate migration in the U.S.

A hybrid smoothing approach is needed for the ACS 2004 data because, unlike the PUMS 5% sample data, the marginal structures also contain irregularities, albeit at lower levels than those found in the origin-destination-specific patterns. For a highly irregular situation such as this, we can use the log-linear with offset approach to incorporate smoothed marginal structures, obtained with model migration schedules. This approach provides a compromise between the more intensive model schedule approach and the unsaturated log-linear approach.

Our theoretical model for smoothing the migration data is the unsaturated model specified in Eq. (4.7). For the ACS 2004 PUMS data, however, the λ_{ix}^{OA} and λ_{jx}^{DA} association terms of the reported data also contain irregularities that would carry forward in the predicted flows. Furthermore, a model without these terms would be considered too simplistic to accurately capture the age-specific migration patterns. Thus, we propose the following log-linear with offset model to smooth the ACS 2004 PUMS data:

$$\ln\left(\hat{n}_{ijx}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \ln\left(n_{ijx}^*\right), \quad i \neq j, \tag{4.8}$$

where the offset, n_{ijx}^*, contains smoothed versions of the association terms, λ_{ix}^{OA} and λ_{jx}^{DA}. The smoothed versions of λ_{ix}^{OA} and λ_{jx}^{DA} were created by fitting Rogers-Castro model migration schedules to the aggregate in-migration and out-migration age compositions, i.e., $N_{i+}(x)$ and $N_{+j}(x)$, which were then divided by A_x(or $N_{++}(x)$) to obtain smoothed versions of OA_{ix} and DA_{jx} multiplicative components (see Chapter 3 for a discussion of the links between multiplicative components and log-linear association terms). The offset, n_{ijx}^*, was then constructed by multiplying the smoothed estimates of OA_{ix} and DA_{jx} by all the other multiplicative components, i.e., T, O_i, D_j, A_x and OD_{ij}, as reported by the ACS 2004.

Some selected results from the hybrid log-linear approach described above are illustrated in Fig. 4.14, along with corresponding model migration schedule and unsaturated log-linear model fits. The Hawaii to Alaska flow represents a situation where only seven data points are available. The Hawaii to California and California to Oregon flows are cases where the patterns are highly irregular. And, the Washington to Oregon flow contains an age profile that is fairly regular with the exception of a small peak at the 50–54 age group. The log-linear with offset model appears to produce the most reasonable results. Clearly the unsaturated model is inappropriate because the irregularities in the marginal structures are carried forward. Model migration schedules have the advantage of making the most use out of the reported data but they involve a large amount of work and can fail when the data are highly irregular, such as these are. For the ACS 2004 PUMS sample data, we were able to fit model schedules to only 12 of the 20 flows in the table. Eight flows were deemed very difficult or impossible to fit model schedules to. These included the Alaska-Hawaii, Alaska-Oregon,

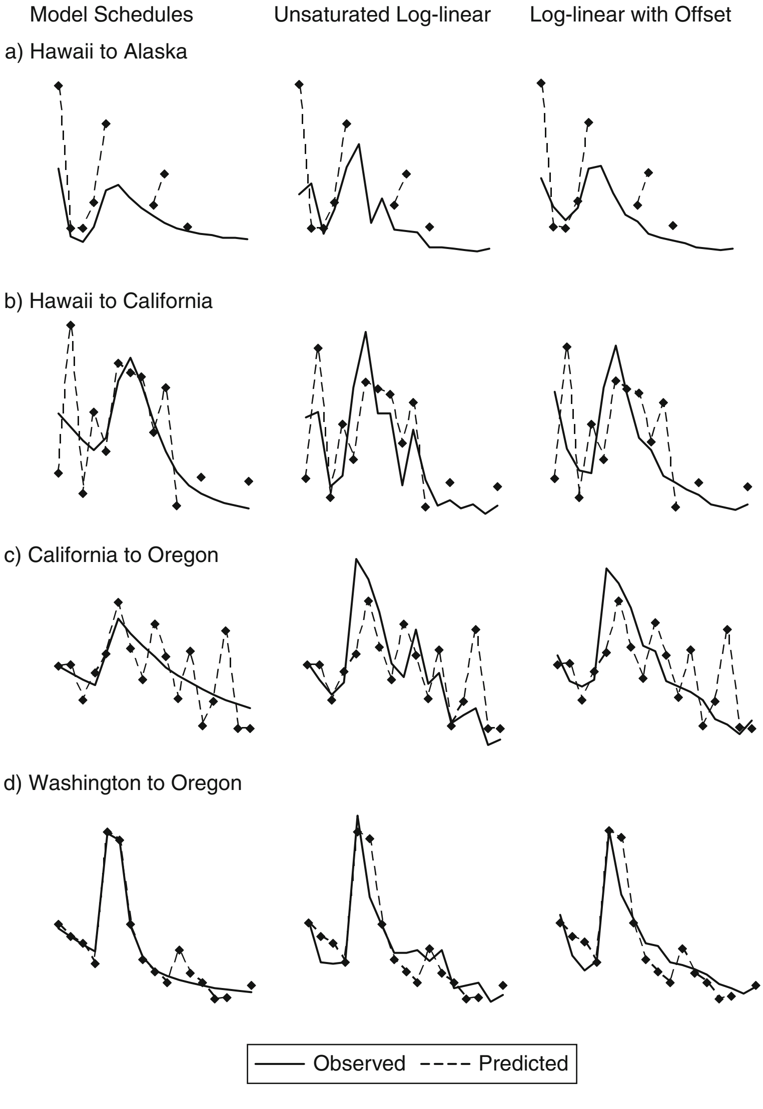


Fig. 4.14 Selected ACS 2004 observed versus predicted age-specific migration flows using Rogers-Castro model migration schedules, the unsaturated log-linear model, and the log-linear with offset model

Alaska-Northeast, Hawaii-Oregon, Oregon-Hawaii, Washington-Hawaii, Northeast-Oregon and Midwest-Oregon flows. Finally, we have examined the corresponding ACS 2005–2007 PUMS sample data and found the data to be of similar poor quality.

4.5.4 Summary

In this section, we have presented several log-linear methods for smoothing a variety of migration data. The focus has been on spatial and age patterns of migration. Of course, these models could be extended to include more factors or to make better use of data available over time, particularly with regard to obtaining more reliable ACS migration data. For example, Eq. (4.7) could be expanded to include time as a main effect:

$$\ln\left(\hat{n}_{ijx}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_t^T + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA}, \tag{4.9}$$

or as interaction terms with origin and destination:

$$\ln\left(\hat{n}_{ijx}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_t^T + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{it}^{OT} + \lambda_{jx}^{DA} + \lambda_{jt}^{DT}. \tag{4.10}$$

Both of these models take advantage of the more reliable structures in the data. The less reliable structures not included in these models would be pooled over time, and should exhibit more regularity in their patterns and in the resulting migration estimates.

4.6 Summary and Discussion

This chapter outlines three basic smoothing techniques that apply three core models: cubic splines, model schedules, and log-linear representations of the data. It begins with a focus on the most straightforward application: the smoothing with cubic splines and then model schedules of the observed relatively large sample (16.67%) migration data reported by the Census 2000 survey. The resulting smoothed data then are taken to be the "gold-standard" against which to compare the corresponding results obtained by applying the very same procedures to the much smaller PUMS 1% sample drawn from the parent population of a 16.67% sample. As expected, the comparison points to the loss of reliability occasioned by the reduction in sample size. But the smoothing of the data from the smaller sample acts to bring the two sets of results surprisingly close to one another, at least for the more populated states.

Turning next to the ACS PUMS sample data, which annually accounts for roughly the same sample size as the PUMS 1% data, we note that some 3 million households received the ACS 2005 questionnaire, giving rise to annual data for about 750 counties with more than 80% of the U.S. population represented (Mather et al., 2005). The ACS PUMS samples are a valuable source of annual socioeconomic data for states and counties. However, analysts studying age- and origin-destination-specific migration flows will be confronted by issues revolving around the relatively small sample sizes associated with high levels of disaggregation and temporal measurement (i.e., the change from a 5-year to a 1-year migration time interval and "averaging" over time). For example, migration between counties with populations less than 65,000 people will be represented by 3- or 5-year averages. Thus, the elimination of the U.S. Census Bureau's long-form questionnaire has made research on migration more complex, and researchers need to exercise caution when using the ACS PUMS samples to analyze migration patterns, particularly those disaggregated by origin, destination, age, sex, or other characteristics.

So what should migration researchers interested in age-specific patterns do in such a situation? They could either pool the samples over several years or somehow correct for the irregularities in the patterns before using them. Pooling data over time has been suggested by the producers of ACS, but that will produce different sets of patterns that may not be as useful to the analyst, and it does not allow the detection of age- and origin-destination-specific changes over time. For ACS PUMS sample data, the best one can do is to identify changes in the more aggregate or stable structures, which our models rely on. Simply pooling the data not only continues the irregular patterns in the annual data (as they are carried forward) but also washes out any differences over time. Smoothing, on the other hand, may provide a way to compare changes in age-specific patterns over time, particularly if the interstate migration *totals* are accurate. The result would be a time series of age-specific migration flow data, which would be very valuable for population projections and analysis.

Smoothing observed data by applying cubic splines and model schedules illustrates an approach that focuses independently on improving each single age profile of migration one at a time. The smoothing process applied to one directional flow does not influence that of another. The third smoothing process described in this chapter differs from the first two in that it adopts an alternative perspective that considers the entire table of interregional migration flows and adopts a smoothing process in which adjustments of a single flow influences that of others. Instead of model schedules we turn to unsaturated log-linear models in which certain interdependencies are ignored. Here the observed data are smoothed by simplified log-linear descriptions in which particular interaction terms, for example three-way interaction terms are removed.

In the event none of the above procedures is effective in improving the data, we must then turn to "repairs" rather than smoothings of the observed data, a topic we take up in the next chapter.

Chapter 5
Imposing Age and Spatial Patterns

5.1 Introduction

In the preceding chapter, we demonstrated methods that are designed to smooth the irregular age patterns of migration which survey data inevitably provide. After fitting model migration schedules, the resulting profiles were more regular and more likely to conform to the expected patterns of age-specific migration. A comparison of the Census 2000 full sample estimates and the PUMS 1% sample estimates showed that the accuracy of the smaller sample estimates also was improved by the smoothing procedures. Figure 4.5 revealed that, on average, smoothing improves the accuracy of migration age profiles. However, the degree of improvement gained by the smoothing procedures is related to the population size of the area being considered. In general, as population size decreases, the average percent error in the migration flows produced from the PUMS 1% sample data increases.

We conclude from our experiments in Chapter 4, which compare the Census 2000 full sample and PUMS 1% sample results, that smoothing techniques should improve the accuracy of the estimates of migration age structure from the Census 2000 PUMS 1% sample and from the ACS data as well. Furthermore, for both the Census 2000 PUMS 1% sample and the ACS PUMS samples, the reliability of the smoothed model schedules diminishes as the population size of the area, i.e., the sample size, decreases. Unlike the Census 2000, there is no way to assess the improvements in reliability due to the smoothing results for the ACS. The methods presented below illustrate ongoing research to develop procedures that can be used to impose more reliable and credible structures on the survey estimates of migration age structure, from surveys such as the ACS.

The methods proposed in this chapter are particularly useful when sample sizes are insufficient to provide reliable age-specific migration flows. These procedures rely both on the survey data in question and on known regularities in migration schedules that have been observed within geographic regions, within families that exhibit similar migration age patterns, and within the same area over time. We build upon the smoothing methods developed in Chapter 4, and go one step further by proposing procedures that are designed especially to alleviate the diminished reliability in the national survey estimates of migration age structure that are associated with the less populated geographic areas. In these situations the survey estimates of

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_5,

migration structure are deemed to be unreliable and still lacking accuracy, even after smoothing techniques have been applied.

As in Chapter 4, the methods are developed and tested by using these data. The results of the methods applied to the Census 2000 PUMS 1% sample data are, once again, systematically compared to the results obtained from the Census 2000 full sample data. This provides a way to gauge the errors and the improvements gained after applying the procedures to the PUMS 1% data. All of the methods presented are based on the principal notion that there are relatively few migration age structures, and that most geographic areas have a migration schedule that conforms to one of these. The strategies are intended to be most useful when applied to less populated areas where they can improve the accuracy and the regularity of the survey-based migration age profiles. The first method is the *regional membership method*, and it is based on the principle that migration age patterns are likely to be similar for areas that are in close geographic proximity. In Section 5.2, we demonstrate the method by considering states that are members of the same U.S. Census Division. A divisional "average" age pattern is derived by consolidating the data for all states in that division. This pattern is then imposed on states within the division that exhibit any major inadequacies in their age patterns.

The second method is the *family membership method*, which is described in Section 5.3. This approach follows Raymer and Rogers (2008) who demonstrated with interstate migration in the U.S. West region that only four age profiles of migration were required to accurately capture the age patterns of migration within the entire system. Migration families offer a parsimonious way of summarizing age structures of migration. In the context of U.S. migration, the survey data, relevant to a particular state, are used to categorize that state as belonging to a migration family, and then the data for all states in a family are combined to define that family's migration age profile. Ultimately, the family membership method provides a migration age structure that can be imposed on each state within the family, and that offers a more reliable alternative to the simple survey-based estimates of a state's migration age structure. Section 5.3.1 describes the steps for determining the migration families from the model migration schedules of the 26 more populated states, which were derived from the PUMS 1% sample data and are considered to be represented accurately by the smoothed data. The model schedules are classified into four families that we believe are able to adequately represent the principal variations in the individual state model schedules.

In Section 5.3.2, each of the 25 less populated states (and the District of Columbia) is assigned to a family based on the parameters of its fitted model schedule. Then the assigned migration family profile is imposed on each state in the family. The success of the family membership method is reported, and it is compared to the regional membership method. The reliability of each method is measured by the degree of alignment between the model migration schedule imposed by the method, and derived from the PUMS 1% sample data, with the corresponding model migration schedule derived from the Census 2000 full sample data.

The regional membership method and the family membership method are expected to improve the ACS estimates of migration schedules for the less populated states. Both of these methods are applied to the ACS 2007 PUMS data, and the results are demonstrated in Section 5.4. Because the ACS is administered yearly, there is an additional method available for the ACS that is not possible with the decennial censuses. This method consolidates the past years of survey data into a temporal "average," and the resulting model migration schedule is used to impose the average migration age structure for the most recent year. This is called the *temporal aggregation method,* and it is developed in Section 5.4.1, where the ACS 2005–2007 PUMS sample data are used to impose the 2007 migration age profile for each of the states.

In Section 5.4.2, the regional membership and the family membership methods are applied to the ACS 2007 data for the more populated states, and the model schedules resulting from the temporal aggregation method are compared with the ACS 2007 derived model schedules. The methods also are contrasted with each other and the relative degrees of success are discussed.

In Section 5.4.3 the methods are applied to the less populated states using the ACS 2007 PUMS sample data, and strategies are presented for assessing the usefulness of one method over another. Section 5.5 is devoted to log-linear methods for imposing spatial migration patterns, and Section 5.6 concludes the chapter with a summary and discussion.

5.2 The Regional Membership Method for Imposing Migration Age Structures

The regional membership method is based on the principle that populations residing in areas that are in close geographic proximity will exhibit similar migration age profiles. If a region is defined so that the subareas contained within it have similar age patterns of migration, then it seems efficient to generate a regional "average" migration age structure that makes use of all survey observations in the region as a whole, and then impose that structure on each of the subareas. Moreover, if a particular subarea has a small population with inadequate migration data, then it seems clear that pooling all of the survey observations in the surrounding region will generate a more reliable estimate than the survey estimate. This is the logic of the regional membership method.

Imposing a migration schedule in this method is a four-step process. First, the region is defined by a collection of subareas that are in close geographic proximity and thought to have similar age patterns of migration. Second, regional "average" model schedules are generated, based on all survey observations in that region. Third, this schedule is scaled up or down to produce schedules that reflect the total aggregate migration propensity (or level) of each subarea in the region. Finally, the adjusted "average" model schedules are imposed on each of the subareas in the region.

We begin our demonstration of the regional membership method with a baseline application of the method which defines a single national region, i.e., one that includes all fifty states and the District of Columbia. We then generate a national model migration schedule, which uses all observations in the Census 2000 PUMS 1% sample data to estimate the national "average" migration age profile based on all interstate migrations between 1995 and 2000. Although we assume, with this demonstration, that all state migration age patterns are similar, we do not assume that all state populations exhibit the same propensities for out-migration. Therefore, before one national migration model schedule can be imposed on a particular state's migration age pattern it must be adjusted by that state's migration propensity. This is accomplished by rescaling the national model migration schedule by the state's aggregate level of out-migration as measured by that state's gross migraproduction rate (*GMR*) as observed in the state's PUMS 1% sample data, thus ensuring that the method for imposing migration structure does not change the total level of out-migration observed in the survey data. (See Section 2.2 for a description of *GMR*.)

Figure 5.1 demonstrates the national model migration schedule after it has been adjusted by the migration levels observed for California, Connecticut, Florida and Wyoming. Each of the four state schedules has the same migration age profile as the national profile, but the levels vary. Wyoming, the least populated and most rural of the four states, shows the highest propensity for out-migration. Connecticut has the next highest, followed by Florida and then California.

Figure 5.2 displays the accuracy of the imposed national model migration schedules for the four selected states, and the larger states, California and Florida, have imposed patterns that are quite similar to the "gold standard" model migration schedules. However, for the less populated states, Connecticut and Wyoming, the imposed schedules are visibly less similar to the schedules derived from the full sample data. The success of this baseline application of the regional membership method is measured by the *MAPE* statistic, which captures the differences between the imposed national migration model schedules and the corresponding migration

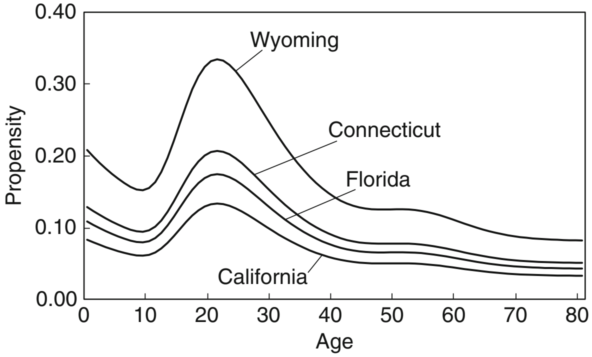


Fig. 5.1 The national model migration schedule after adjustments for each state's *GMR*

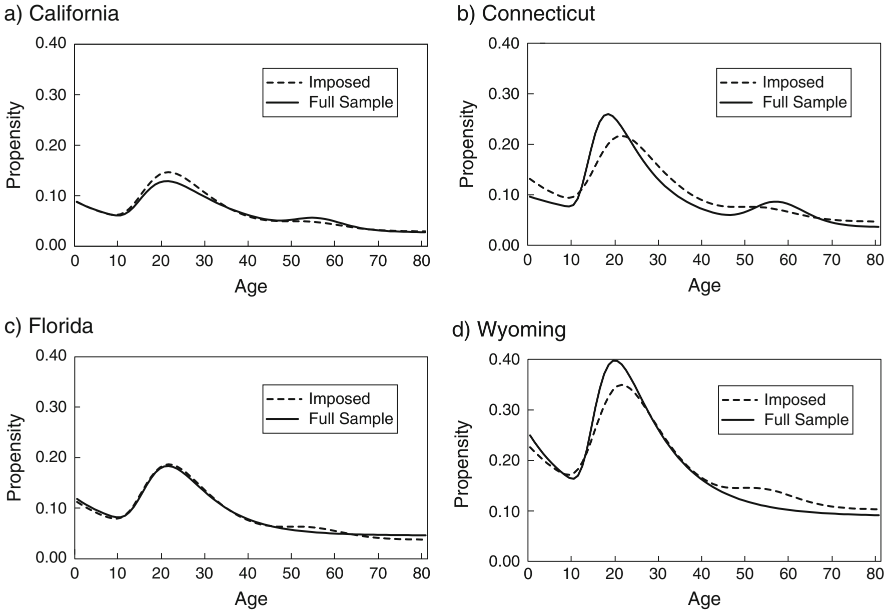


Fig. 5.2 The imposed national model schedule as compared to the model schedules derived from the full sample data

model schedules derived from the full sample data. (See Chapter 4 Section 4.2 for a definition of *MAPE*.)

The *MAPE*s for all states are reported in Table 5.1. The *MAPE*s that resulted from imposing the national model schedule on California and Florida are 5.52% and 5.83%, respectively, and the *MAPE*s for Connecticut and Wyoming are 17.05% and 14.54%, respectively, thereby supporting the visual similarities and differences displayed in Fig. 5.2.

The national model migration schedule provides a simplistic example of the regional membership method, but it serves as a baseline for comparison with a more realistic application of the method. For example, consider regions that are defined as the nine U.S. Census divisions. For this illustration, the state out-migration data obtained from the PUMS 1% sample data were consolidated to the division level then splined and fitted by model schedules as described in Chapter 4, resulting in division "average" model migration schedules. Before imposing these division schedules on the states within the divisions, they were rescaled so that each state's imposed model migration schedule had the same *GMR* that was observed in the PUMS 1% sample. Figure 5.3 displays, for four of the nine divisions, the "average" migration schedules, after adjusting for each member state's *GMR*. The New England model schedule exhibits a pronounced labor peak, a retirement peak, and relatively low levels of infant migration. The Pacific model schedule is similar to the New England model schedule, except that it has higher infant migration levels.

Table 5.1 A comparison of the *MAPE*s generated from two applications of the regional membership method

U.S. Census Division	State	Imposed national model schedule *MAPE*	Imposed division model schedule *MAPE*
Northeast	**Connecticut**	17.05	5.77
	Maine	20.52	15.03
	Massachusetts	18.20	7.54
	New Hampshire	16.65	11.17
	Rhode Island	16.40	12.96
	Vermont	22.31	14.44
	Avg *MAPE*	17.76	10.49
N Atlantic	New Jersey	13.55	7.34
	New York	10.44	5.91
	Pennsylvania	12.67	11.42
	Avg *MAPE*	12.22	8.23
Central	Illinois	5.85	5.74
	Indiana	7.58	4.63
	Michigan	14.48	8.36
	Ohio	9.27	8.72
	Wisconsin	15.00	12.92
	Avg *MAPE*	10.44	8.07
Midwest	**Iowa**	15.85	10.07
	Kansas	17.37	7.38
	Minnesota	12.83	12.65
	Missouri	9.27	5.59
	Nebraska	19.68	7.38
	North Dakota	36.20	21.47
	South Dakota	22.74	10.38
	Avg *MAPE*	19.13	10.70
M Atlantic	**Delaware**	9.44	8.62
	DC	18.76	11.85
	Florida	5.83	8.40
	Georgia	9.66	2.51
	Maryland	8.49	6.07
	North Carolina	18.66	10.06
	South Carolina	15.23	6.90
	Virginia	17.42	8.04
	West Virginia	12.69	7.57
	Avg *MAPE*	12.91	7.78
South	Alabama	15.21	4.37
	Kentucky	10.21	5.02
	Mississippi	13.53	4.81
	Tennessee	13.97	3.14
	Avg *MAPE*	13.23	4.34
S Central	**Arkansas**	6.66	9.95
	Louisiana	19.24	10.69
	Oklahoma	17.82	10.45
	Texas	10.44	2.57
	Avg *MAPE*	13.54	8.41

Table 5.1 (continued)

U.S. Census Division	State	Imposed national model schedule *MAPE*	Imposed division model schedule *MAPE*
Mountain	Arizona	8.39	7.06
	Colorado	5.41	1.47
	Idaho	12.60	12.46
	Montana	19.63	21.10
	Nevada	12.70	13.85
	New Mexico	5.17	4.78
	Utah	12.97	11.81
	Wyoming	14.54	14.84
	Avg *MAPE*	11.43	10.92
Pacific	**Alaska**	8.52	7.98
	California	5.52	3.70
	Hawaii	42.68	45.56
	Oregon	8.32	9.92
	Washington	6.91	8.94
	Avg *MAPE*	14.39	15.22

Note: States in boldface are the 25 less populated

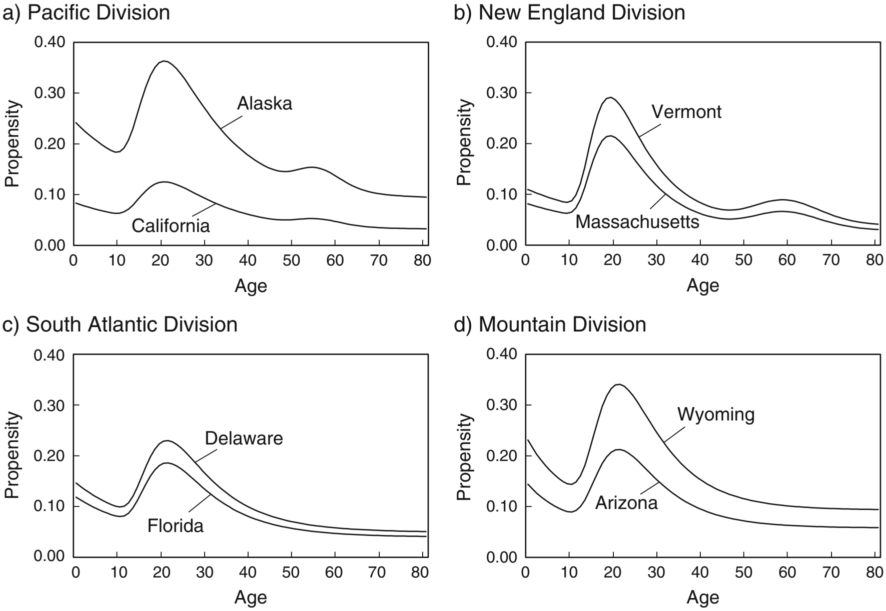


Fig. 5.3 The imposed division model schedules for the states with the highest and lowest migration propensities within four U.S. Census Divisions

The Mountain and South Atlantic schedules both have a pronounced labor peak, relatively high infant migration, and no retirement peak.

Within each division there is variation in the propensity for migration, and the extent of variation is exhibited in Fig. 5.3. The imposed model schedules are displayed for the states with the highest and the lowest migration levels within the four selected divisions. In all four divisions, the less populated states had higher observed *GMR*s than the more populated states, and this is confirmed by their higher levels of migration as displayed in Fig. 5.3. Within the Pacific Division there is more variation than in all of the other divisions, and the imposed profile for Alaska has the highest total migration level and California has the lowest. The Mountain Division and the New England Division each exhibit significant variations in migration levels. Wyoming, the least populated state, has the highest migration level, as compared to Arizona which has the lowest migration level in the Mountain Division, and to Vermont which has the highest migration level in the New England Division, as compared with Massachusetts, which has the lowest. The imposed schedules for the states in the South Atlantic Division have relatively little variation in migration levels. Delaware has the highest level of migration, and it is just slightly higher than schedule imposed on Florida which has the lowest migration propensity. (The District of Columbia is not considered here because it is not a state, and it has a level of migration that is distinctly different than all states.)

In Fig. 5.4 the imposed division model migration schedules are visually contrasted with the schedules derived from the full sample data for the selected

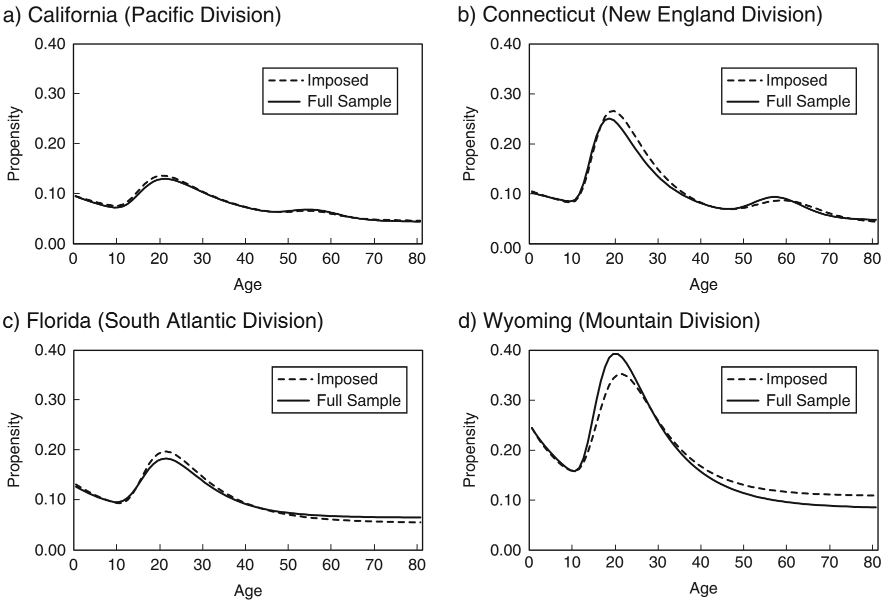


Fig. 5.4 The imposed division model schedules as compared to the model schedules derived from the full sample data

states. For California (Pacific), for Florida (South Atlantic), and for Connecticut (New England), the method works very well. For these three states the imposed division model schedules coincide almost perfectly with the model schedule derived from the full sample data. For Wyoming (Mountain), the imposed model schedule exhibits a lower peak and a higher propensity for migration in the retirement years than the state's model schedule derived from the full sample data.

The differences and the similarities that are visually apparent in Fig. 5.4 are summarized with *MAPE* statistics in Table 5.1. In addition, the accuracy of the "national" versus the "division" application of the regional membership method is contrasted in Table 5.1. For example, for California, the difference between the model schedule imposed by the division and the full sample data is reflected by the *MAPE* = 3.70, which is only a slight improvement over imposing the national schedule on the state of California (*MAPE* = 5.52). For Connecticut, the division model schedule (*MAPE* = 5.77) is much more accurate than the national model schedule (*MAPE* = 17.05). For Florida, on the other hand, the national model schedule was slightly more accurate (*MAPE* = 8.40) than the division model schedule (*MAPE* = 5.83). Both applications of the regional membership method were accurate to a similar degree for Wyoming (national *MAPE* = 14.54, divisional *MAPE* = 14.84). In general, the strategy of imposing the national model schedule works well for the most populated states such as California (*MAPE* = 5.52), New York (*MAPE* = 10.44), Illinois (*MAPE* = 5.85), and Florida (*MAPE* = 5.83), while the strategy of imposing the division model schedule works well when the states within the division are relatively homogeneous, as in the case of the Middle Atlantic (mean *MAPE* = 8.23), the East North Central (mean *MAPE* = 8.07), and the East South Central (mean *MAPE* = 4.34).

5.3 The Family Membership Method for Imposing Migration Age Structures

Little and Rogers (2007) showed that the model schedules representing the age compositions of out-migrants for three levels of U.S. geography (states and consolidated metropolitan areas, metropolitan statistical areas, and counties) could be reduced to just a few families of schedules for each level of geography (see also Raymer & Rogers, 2008). In this section, we assign geographic units into migration families that tend to exhibit similar age profiles of migration. Based on the model schedules that were fitted to the PUMS 1% sample data (as discussed in Section 4.3), the primary family defining characteristic is the presence or absence of a retirement peak. At the same time, we expect that members of the same family will have varying levels of migration. Therefore, areas should be assigned to a family according to their *profile* parameters and not their *level* parameters. (Section 2.2.2 sets out the analytic form of the Rogers-Castro model migration schedule and a description of each component of the model.)

Within the families that have or do not have a retirement peak, child dominant and labor dominant families are distinguished by a comparison of the relative migration propensities for infant migration and for the labor force peak ages. The ratio of the *labor force peak value* (evaluated as $a_2 + \exp(-20\alpha_1) + c$) to the *infant migration peak value* (evaluated as $a_1 + c$) indicates whether the flow is either a labor dominant or a child dominant flow. Higher ratios are indicators of a labor dominant migration family, and lower ratios are associated with a child dominant migration family.

As a result of this classification scheme, there are four possible migration families: (1) retirement peak, labor dominant; (2) retirement peak, child dominant; (3) no retirement peak, labor dominant; and (4) no retirement peak, child dominant. Once the areas are assigned to one of these four migration families, certain components of the family's flow schedule are imposed on each of the family members. For example, the labor force profile parameters (α_1, α_2, μ_2, λ_2), averaged over the family members, are assigned to the imposing family model schedule. In addition, if there is a retirement peak, the family average of each of the retirement peak parameters (a_3, α_3, μ_3, λ_3) is assigned to the imposing model schedule. The *level* parameters (a_1, $a_{2,}$ and c) are not assigned by the family. Instead, they are drawn from the member's PUMS 1% sample data or from its corresponding model schedule.

5.3.1 Defining Families of Out-Migration Flows

In Section 4.3, we found that the smoothed age profiles of migration that resulted from applying model schedules to the splined PUMS 1% sample data provided a good representation of the observed values for the 26 states with populations over four million persons. In this subsection, we use the parameters from these model schedule fits to identify model schedule families. This was carried out in the following way. First, the model schedules exhibiting a retirement peak were separated from those not exhibiting them. Second, all flows were classified as being labor- or child-dominant based on the ratios of labor force peak values to the initial infant migration peak values. The groups were made based on natural breaks found in the distribution of ratios.

Among the 26 most populated states, there were eight with retirement peaks. Of those, two were classified in the child dominant family (California and Illinois) and six in the labor dominant family (Indiana, Massachusetts, Michigan, Minnesota, New Jersey, and New York). Of the remaining 18 states without a retirement peak, 11 were classified in the child dominant family (Arizona, Florida, Georgia, Louisiana, Maryland, North Carolina, Ohio, South Carolina, Tennessee, Washington, and Wisconsin) and seven in the labor dominant family (Alabama, Colorado, Kentucky, Mississippi, Pennsylvania, Texas, and Virginia).

Once the family classifications were assigned, the steps for generating the imposing model schedule to represent each family were as follows. Over all members in a family, we averaged each of the labor force "pure" *profile* parameters (α_1, α_2, μ_2, λ_2) , and, if there was a retirement peak, we took the average of each of the retirement peak parameters (a_3, α_3, μ_3, λ_3). These averaged parameters were then assigned to the imposing family model schedules. Because we expected members of the same family would have varying levels of migration, the *level* parameters (a_1, a_2, and c) were not assigned by the family. Instead, a_1 and a_2 were preserved from the state's model schedule, and the c parameter was calculated as the average propensity to migrate over age 60 as indicated by the splined age profiles derived from PUMS 1% sample data (discussed in Section 4.3). As in the regional membership method, the final step in the family membership method was to rescale the schedule imposed by the method, in a way that preserves the *GMR* that was observed in the PUMS 1% sample.

Figure 5.5 shows the imposed migration schedules that resulted from the family membership method for two states selected from each of the four families. It demonstrates the similarities in migration age profiles within a family, and, at the same time, it reveals that the height of the infant migration peak, the labor peak and the retirement peak can be quite different within a particular family. The close correspondence between the schedules estimated from the full sample data and those imposed by the family membership method is visually demonstrated for California, Massachusetts, Florida, and Pennsylvania in Fig. 5.6.

The *MAPE* statistics are reported in Table 5.2 for the 26 large states grouped into four families. The average *MAPE* of 7.80 suggests that the family membership method is an effective option for indirectly estimating the migration age profiles for the most populated states. Only three states (Minnesota, Wisconsin, and Texas) had

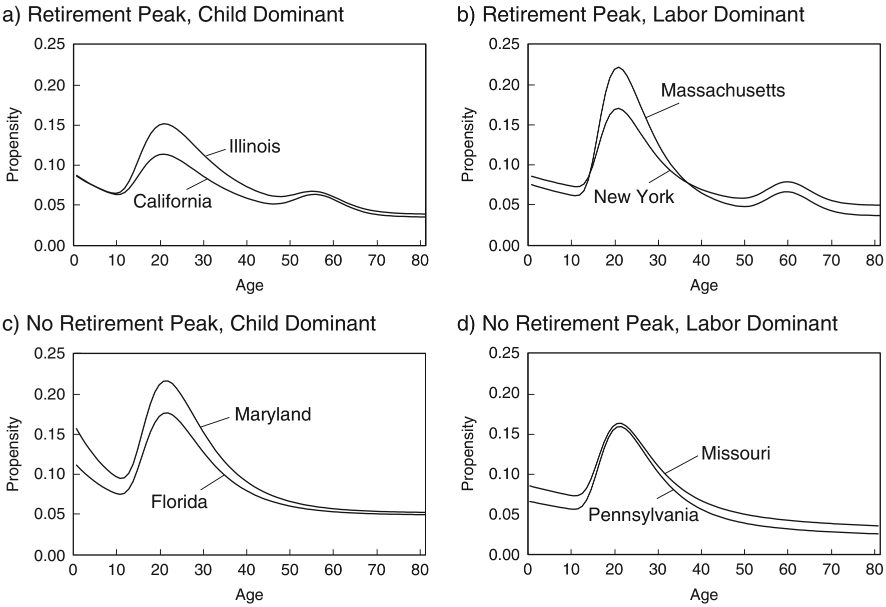


Fig. 5.5 The imposed family model schedules for selected more populated states

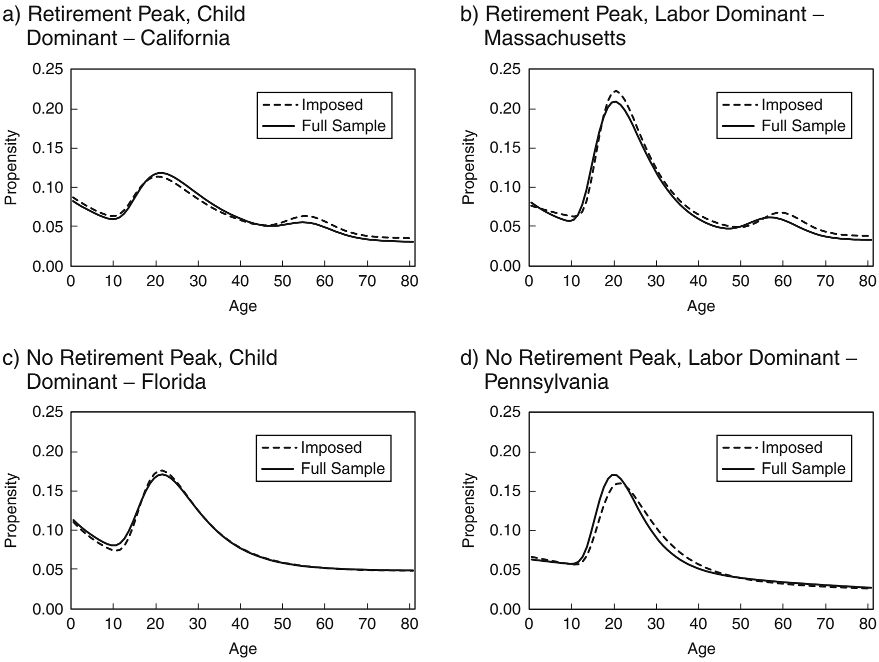


Fig. 5.6 The imposed family model schedules as compared to the model schedules derived from the full sample data for selected more populated states

migration schedules that did not conform well to their migration family, and their *MAPE* scores were 15.09, 16.39, and 16.41, respectively. The average of the *MAPE* scores for the family membership method (7.80) is comparable to the average for the regional membership method (6.91). For 11 of the 26 states, the family membership method actually performed better than the regional membership method, and for 15 states the regional membership method was superior.

5.3.2 *The Family Membership Method Applied to the Less Populated States*

Although the smoothed profiles from the PUMS 1% sample data provide good representations of the true profiles for the states with populations over 4 million, they generally are not as good for states with less than 4 million people. For these states, and for other areas where survey data may be insufficient, we may apply the family membership method, which uses characteristics of the more realistic schedules borrowed from the most populated areas, and yet preserves some of the unique character of the model schedules estimated from the 1% PUMS sample data for the less populated areas. It is reasonable to conjecture that those states with fewer than 4 million

Table 5.2 A comparison of the family membership method and the regional membership method (imposing the division model schedule) for the more populated states

	State classification	Family membership method *MAPE*	Regional Membership Method *MAPE*
Retirement peak:			
Child dominant	California	7.97	3.70
	Illinois	4.19	5.74
Labor dominant	Indiana	9.61	4.63
	Massachusetts	9.10	7.54
	Michigan	7.20	8.36
	Minnesota	15.09	12.65
	New Jersey	8.07	7.34
	New York	5.25	5.91
No retirement peak:			
Child dominant	Arizona	4.86	7.06
	Florida	1.71	8.40
	Georgia	9.00	2.51
	Louisiana	7.60	10.69
	Maryland	7.25	6.07
	North Carolina	2.03	10.06
	Ohio	5.68	8.72
	South Carolina	6.06	6.90
	Tennessee	8.16	3.14
	Washington	5.18	8.94
	Wisconsin	16.39	12.92
Labor dominant	Alabama	9.55	4.37
	Colorado	4.20	1.47
	Kentucky	6.46	5.02
	Missouri	6.36	5.59
	Pennsylvania	6.61	11.42
	Texas	16.41	2.57
	Virginia	12.68	8.04
	Overall Average M*APE*	7.60	6.87

persons should have out-migration flows that are similar in age profile to those states with more than 4 million persons. Thus, we assign the small states to the appropriate family using their model schedule parameters. If these indicate there is a retirement peak, then they are assigned to the retirement peak family. As before, the ratio of the labor force peak value (as reflected by $a_2+\exp(-20\alpha_1)+c$) to the initial infant migration peak value (as reflected by $a_1 + c$) indicates whether the flow is either a labor dominant or a child dominant flow (Rogers & Castro, 1986).

To calculate the imposed schedules we take the averages of the labor force *profile* parameters and all retirement peak parameters from the large states (by family), and then impose these parameters onto the flows from the small states (by family). For each small state flow, we use the state's observed level parameters (a_1, a_2) and the c parameter, which is calculated as the average propensity to migrate over age 60 that is indicated by the state's splined data. Finally, we impose the observed gross-migraproduction rate (*GMR*) by rescaling the fitted data, thus ensuring that no changes arise in each state's level of out-migration.

Only Connecticut was found to have a retirement peak, and it was placed in the labor dominant family with a retirement peak. Most of the less populated states (16 of them) were classified as members of the child dominant family, and 8 were categorized as being in the labor dominant family. None of the less populated states fell into the "retirement peak, child dominant" family. Figure 5.7 demonstrates the similarities in the shapes of the imposed schedules within each family, and, at the same time, it shows that there are variations in the levels of the imposed schedules for representative states within each family. Figure 5.8 visually demonstrates the closeness between the imposed model schedules and the corresponding full sample schedules for selected states within each family.

Table 5.3 displays the classification into families for the 25 least populated states and the *MAPE* statistics associated with the results of the family membership method and the regional membership method for comparison. When the imposed schedules from the family membership method are compared to those of the smoothed schedules derived from the full sample data, on average the *MAPE* is 11.48 for the least populated states. Only the District of Columbia, in the child dominant family, had a *MAPE* above 15.00. In contrast, one half of the states in the labor dominant family had *MAPE*s above 15.00, which suggests that the labor dominant migration family is not as homogeneous as is the child dominant migration. When

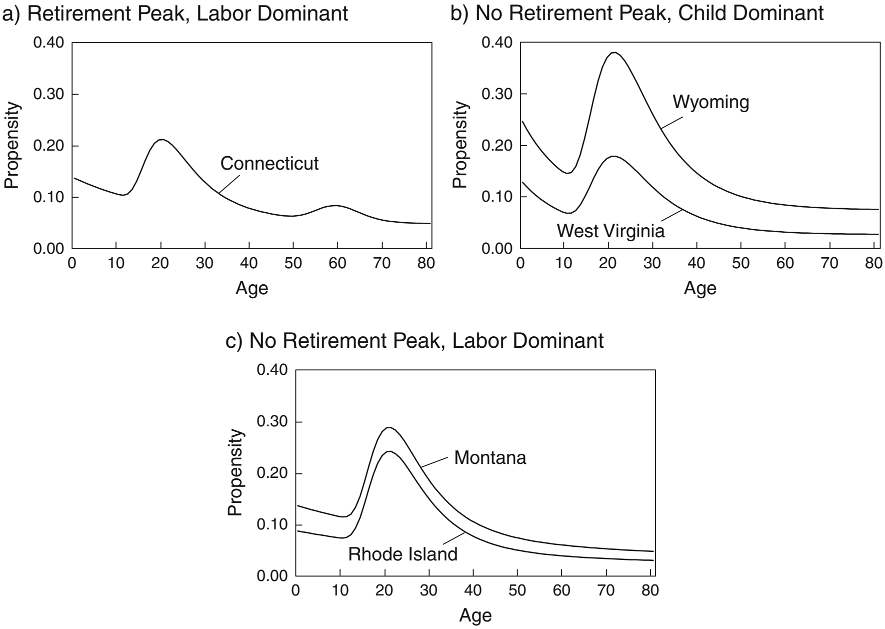
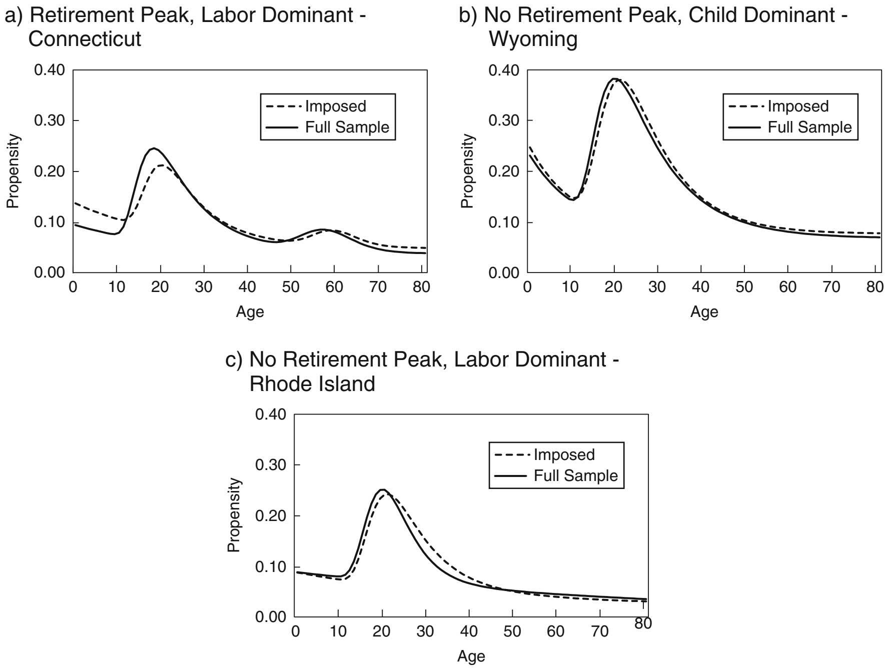


Fig. 5.7 The imposed family model schedules for selected less populated states


a) Retirement Peak, Labor Dominant - Connecticut

b) No Retirement Peak, Child Dominant - Wyoming

c) No Retirement Peak, Labor Dominant - Rhode Island

Fig. 5.8 The imposed family model schedules as compared to the model schedules derived from the full sample data for selected less populated states

the *MAPE*s from the family membership method and from the regional membership method are compared it seems that, on average, the two methods perform equally well. The average *MAPE* resulting from the regional membership method is 12.46 compared to 11.48 for the family membership method. Among the less populated states, 13 of 25 had a smaller *MAPE* for the family membership method than for the regional membership method. In addition, 12 of the 25 states had 10% or less error from the family membership method, and 13 of the 25 states had 10% or less error from the regional membership method. However, if the best method is chosen to impose the migration age structure for the less populated states, the average *MAPE* improves from 11.48 to 8.61, which suggests there are advantages in having two methods to choose from.

In general, each method worked well under certain conditions. The regional membership method worked well, especially when the U.S. Census divisions were used to define the regions, and when the states within a division were homogeneous in their migration age structures. For example, Mississippi is a less populated state in East South Central, where all four states have similar migration schedules. Here, the model schedule imposed by the regional membership method was very similar to its model schedule derived from its full sample data ($MAPE = 4.81$). New Mexico is another example of a less populated state where the regional membership method was successful ($MAPE = 4.78$). The states within the Mountain Division are

Table 5.3 A comparison of the family membership method and the regional membership method (imposing the division model schedule) for the less populated states

	State classification	Family membership method *MAPE*	Regional membership method *MAPE*
Retirement peak:			
Labor dominant	Connecticut	14.97	5.77
No Retirement peak:			
Child dominant	Alaska	14.13	7.98
	Arkansas	4.47	9.95
	District of Columbia	16.99	11.85
	Hawaii	14.48	45.56
	Idaho	8.3	12.46
	Iowa	10.15	10.07
	Kansas	5.67	7.38
	Nevada	3.21	13.85
	New Hampshire	14.99	11.17
	New Mexico	3.58	4.78
	Oklahoma	5.59	10.45
	Oregon	6.94	9.92
	South Dakota	14.1	10.38
	Utah	8.56	11.81
	West Virginia	9.77	7.57
	Wyoming	6.25	14.84
Labor dominant	Delaware	11.02	8.62
	Maine	16.92	15.03
	Mississippi	23.23	4.81
	Montana	8.75	21.10
	Nebraska	21.79	7.38
	North Dakota	13.4	21.47
	Rhode Island	11.03	12.96
	Vermont	18.79	14.44
	Overall Average *MAPE*	11.18	12.38

not homogeneous, but New Mexico happened to have a migration age structure that aligned closely with the division's average age structure, which is shaped in large part by two states that are neighbors to New Mexico and are the most populated states in the division (Arizona and Colorado).

The family membership method worked notably well for three states within the heterogeneous Mountain Division. In the child dominant migration family (with no retirement peak), the imposed model schedule corresponded well with the full sample derived model schedule for Idaho ($MAPE = 8.30$), Nevada ($MAPE = 3.21$), and Wyoming ($MAPE = 6.25$). In contrast, the *MAPE*s associated with the regional membership method were 12.46 for Idaho, 13.85 for Nevada, and 14.84 for Wyoming. Similarly, in the labor dominant migration family (with no retirement peak), the family membership method performed well for Montana ($MAPE = 8.75$) in comparison to the regional membership method ($MAPE = 21.10$).

5.4 Imposing Migration Age Structures with ACS Data

In Sections 5.2 and 5.3, the regional membership and the family membership method for imposing migration age structures were developed using the Census 2000 PUMS 1% sample data, and the accuracy and the reliability of the imposed model schedules were compared to those derived from the Census 2000 full sample data. As expected, the imposed model schedules corresponded well with the full sample derived model schedules in the most populated states, and the average *MAPE*, over the 26 largest states, was 6.91 for the regional membership method and 7.80 for the family membership method. Our analysis focused primarily on the least populated states, where the PUMS 1% sample data results are less reliable, and where there is clearly more need for methods that can impose more reliable migration age structures.

In this section, we demonstrate that the proposed methods can be applied to the ACS PUMS sample data to improve the estimates of migration age structures. The ACS offers a special challenge, over and above the decennial Census 2000, in that it has no full sample data source that can be used to verify the reliabilities of the methods for imposing migration age structures. To get around this shortcoming, we use the model schedules derived from the ACS PUMS sample data for the more populated states as the "gold standard" and assume that these estimates are quite accurate and can be viewed as the "true" migration age structures for those states.

At the same time, the ACS has the advantage of being administered annually. Since 2007 is the third year of full implementation of the ACS, there are three years of ACS data (2005, 2006, and 2007) that can be combined to produce a set of survey observations that is three times larger than the number of observations in the ACS 2007 survey alone. Using all three years of data (2005–2007) to derive a model schedule that can be imposed on the 2007 migration age structures provides an additional method which can be applied uniquely to the annual ACS PUMS sample data, and it is called the *temporal aggregation method.*

5.4.1 The Temporal Aggregation Method for Imposing Migration Age Structures

At the inception of the ACS, the U.S. Census Bureau recognized there would be problems with sampling error, especially for the smaller (less populated) areas, and they have argued that the 3-year and 5-year aggregated estimates should be used instead of the single-year estimates for these areas. In addition, the Census Bureau will publish these estimates every year for the most common tables. The temporal aggregation method is equivalent to these 3-year ACS estimates, but it is tailored specifically for application to migration age structures. The temporal aggregation method, can be implemented by any user of the ACS PUMS sample data by simply combining all of the observations available in each of the three ACS PUMS files. In this case, we are estimating state to state migration between 2006 and 2007, but we are using the ACS 2007 as well as the ACS 2005 and 2006 PUMS sample data, and we simply assume that the 2006 and the ACS 2005 data provide "proxy" estimates of migration between 2006 and 2007.

The temporal aggregation method is based on the assumption that migration age structures are relatively stable over a 3-year period, and it assumes that the combination of the observations for the three most recent annual surveys will yield reasonably accurate estimates of the most recent annual migration age structures. To test these assumptions we selected four of the largest states (California, New York, Florida, and Texas) and derived the model schedules from the annual ACS PUMS sample data. Since these are the largest states, their ACS derived annual model schedules will have minimal sampling error, and, for our purposes here, we assume they accurately reflect the annual migration structures of the population.

Figure 5.9 shows that the variation in model schedules for 2005, 2006, and 2007 is quite substantial for the most populated states. For each of the four states in Fig. 5.9, the migration propensities change over time, and the most consistent difference between the 2007 model schedules and the 2005 and 2006 model schedules is at the age of the lowest propensity for child migration. This is around age 10 for all four states in 2007, around age 12 for 2006, and around age 15 for 2005. Figure 5.10 displays the model schedules that are imposed by temporally aggregating the ACS data, and these are contrasted with the ACS 2007 derived model schedules, showing

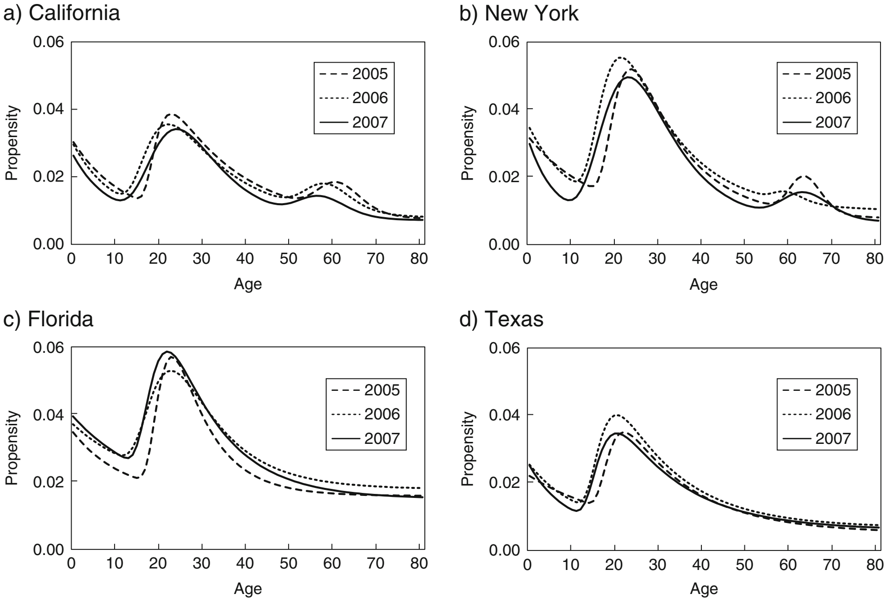


Fig. 5.9 Model migration schedules derived from the ACS 2005, 2006, and 2007 PUMS sample data for selected more populated states

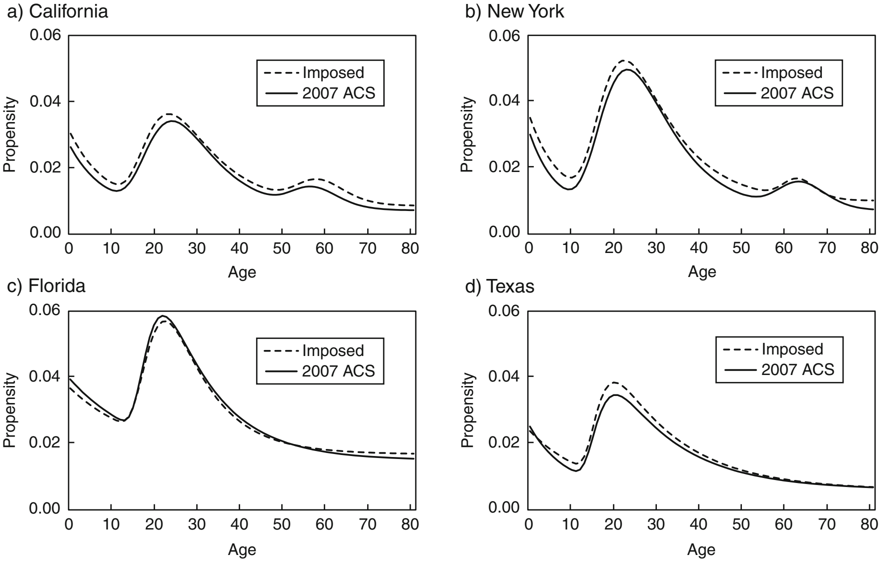


Fig. 5.10 The imposed temporally aggregated (2005–2007) model migration schedules compared to the ACS 2007 derived model migration schedules for selected more populated states

clearly that there are differences between the imposed 3-year "average" schedules and the 2007 annual schedule. The differences are primarily in the migration levels, but, overall, the profiles of the imposed schedules are quite similar to the 2007 annual schedules.

Figures 5.11 and 5.12 reflect a similar dynamic for the less populated states selected (Connecticut, West Virginia, Wyoming, and Nebraska). Again the migration levels change over time, and there are differences between the 2007 schedules and the 2005 and 2006 schedules in the age at the lowest propensity for child migration, occurring around age 10 for the 2007 schedules and later for the other years, especially in the 2005 model schedules. For Connecticut, West Virginia, and Nebraska, labor migration peaks occur at an earlier age in 2007 than in previous years. Nevertheless, when the ACS data for 2005, 2006, 2007 are combined, the resulting model schedules, imposed by the temporal aggregation method, are aligned very closely with the ACS 2007 model schedules, as is visually apparent in Fig. 5.12.

The comparability between the temporally aggregated schedules and the annual ACS 2007 derived schedules is summarized in Table 5.4. The *MAPE* statistics in the first column can be interpreted as an assessment of the discrepancies due to the temporal aggregation method, since it is assumed that the more populated states have ACS 2007 derived model schedules that accurately reflect their migrating populations. Overall, the more populated states have an average *MAPE* of 10.90. Louisiana is a clear outlier, however, due to the unusual migration patterns associated with Hurricane Katrina. If Louisiana is omitted, the adjusted average *MAPE* is 8.73.

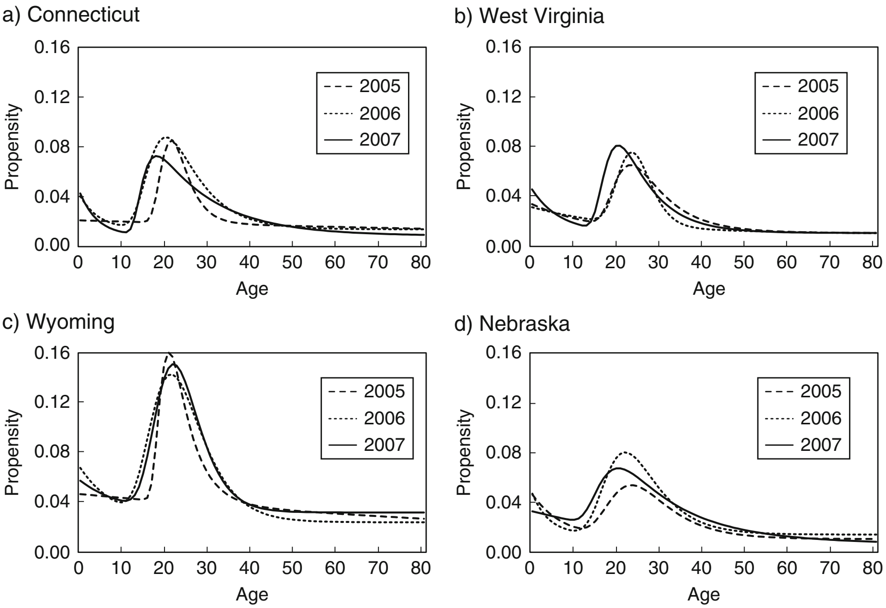


Fig. 5.11 Model migration schedules derived from the ACS 2005, 2006, and 2007 PUMS sample data for selected less populated states

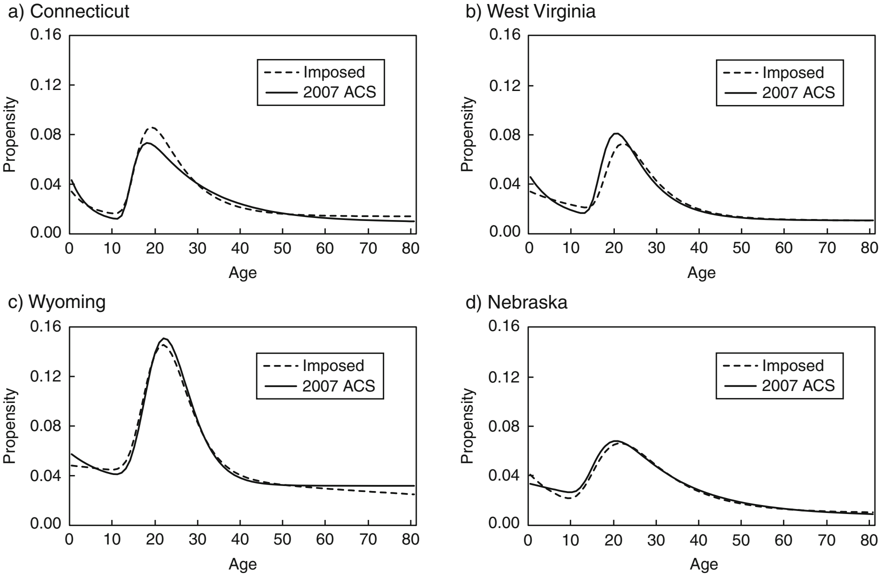


Fig. 5.12 The imposed temporally aggregated (2005–2007) model migration schedules compared to the ACS 2007 derived model migration schedules for selected less populated states

Table 5.4 Goodness-of-fit statistics (*MAPE*) contrasting the model schedules imposed by the temporal aggregation method and the model schedules derived from the ACS 2007

More populated states	*MAPE*	Less populated states	*MAPE*
Alabama	7.38	Alaska	8.64
Arizona	6.42	Arkansas	9.10
California	14.33	Connecticut	15.36
Colorado	5.77	Delaware	16.36
Florida	3.91	District of Columbia	7.72
Georgia	3.77	Hawaii	9.65
Illinois	6.54	Idaho	25.71
Indiana	10.01	Iowa	24.67
Kentucky	6.08	Kansas	8.36
Louisiana	71.09	Maine	17.57
Maryland	10.21	Mississippi	18.05
Massachusetts	9.51	Montana	12.15
Michigan	12.08	Nebraska	6.05
Minnesota	4.69	Nevada	6.69
Missouri	4.04	New Hampshire	9.68
New Jersey	11.67	New Mexico	7.77
New York	12.90	North Dakota	18.17
North Carolina	7.89	Oklahoma	17.51
Ohio	10.64	Oregon	11.88
Pennsylvania	3.38	Rhode Island	13.86
South Carolina	16.96	South Dakota	15.58
Tennessee	6.86	Utah	8.14
Texas	7.29	Vermont	11.43
Virginia	5.07	West Virginia	7.73
Washington	13.19	Wyoming	7.15
Wisconsin	11.78		
Average *MAPE*	8.49*	Average *MAPE*	12.60

**Note*: Louisiana was omitted from this calculation

Of the states with results presented visually in Fig. 5.10, California has the largest *MAPE* (14.33). This relatively high *MAPE* is due to the consistent lift of the imposed schedule above the 2007 schedule. New York has the next highest *MAPE* (12.90), and it is slightly lower than the *MAPE* for California because the two New York schedules come closer together in the older ages. Texas and Florida show *MAPE*s (7.29 and 3.91, respectively) that are consistent with their visual displays in Fig. 5.10.

The *MAPE* results for the less populated states are reported in the right column of Table 5.4. They are not directly comparable to the *MAPE*s for the more populated states, because the model schedules derived from the annual ACS data, for the less populated states, are the result of more sampling error than those of the more populated states, and it would be a mistake to assume these schedules represent the "true" migration age structure of the population for that year. Nevertheless, the results reported for the less populated states, in Table 5.4, are valuable because they quantify the degree of alignment between the imposed schedules and the ACS 2007 schedules. On average, the *MAPE* is 12.76, which suggests that the discrepancies between the annual model schedules and those imposed by the temporal aggregation method are larger for the less populated states than for the more populated states (average *MAPE* = 8.73).

Where there is close correspondence between the imposed schedule and the 2007 annual model schedule, and the *MAPE* is small, say less than 10%, this result suggests there was stability in that state's migration age structure for the years 2005 to 2007, and it gives validity to the imposed schedule as a reliable estimate of the migration age structure for that state. On the other hand, when the *MAPE* is large, as in the case of Idaho (*MAPE* = 25.71), it may reflect there is variation in the migration age structure over the three year period, and the temporal aggregation method is inappropriate, or it may reflect that sampling error has distorted the model schedule derived from the ACS 2007 data.

5.4.2 The Imposing Methods Applied to the More Populated States

We now apply the regional membership and the family membership methods, in addition to the temporal aggregation method, to the ACS 2007 data, targeting the 26 more populated states. The result is three sets of model schedules that are contrasted with the model schedules estimated directly from the ACS 2007 data. In this section we quantify the success of each of the methods because, for the more populated states, we assume that the ACS 2007 derived model schedules are measured without error, and that they provide the "true" migration age structures of the populations of the more populated states. Therefore, the "best" method for imposing the migration age structure is the one with the closest correspondence to the ACS 2007 model schedule.

The regional membership method was applied to the ACS 2007 data as described in Section 5.2, using U.S. Census Divisions as the regions, and the family membership method was applied as described in Section 5.3. However, the family classifications based on the ACS 2007 data are different than they were for the Census 2000 1% PUMS data. With the ACS 2007 data, only three families were identified, in contrast to the four families identified in the Census 2000 1% PUMS data. The state classifications in migration families are reported in Table 5.5. Some states were identified as having a retirement peak, but further distinction into labor or child dominant families was not possible. Of these 6 states, the regional membership method performed better than the other two methods with an average *MAPE* of 8.30. However, two states (Illinois and New Jersey) had slightly better results from the temporal aggregation method, and no states had better results from the family membership method.

For the two families without a retirement peak, i.e., the child dominant and the labor dominant families, the temporal aggregation method proved to be superior to the other methods, having the smallest *MAPE*s on average (6.94 and 8.59, respectively). Over all of these states, on average, the temporal aggregation method

Table 5.5 A comparison of the three imposing methods applied to the more populated states: goodness-of-fit statistics (*MAPE*) contrasting the method's model schedule with the model schedule derived from the ACS 2007

	State classification	Family membership method *MAPE*	Regional membership method *MAPE*	Temporal aggregation method *MAPE*
Retirement peak:				
	California	11.08	5.73	14.33
	Illinois	11.17	8.88	6.54
	Massachusetts	22.86	8.50	9.51
	New Jersey	12.94	11.89	11.67
	New York	11.76	8.04	12.90
	Ohio	7.73	6.75	10.64
	Average *MAPE*	12.92	8.30	10.93
No retirement peak:				
Child Dominant	Arizona	7.77	4.48	6.42
	Georgia	13.70	6.15	3.77
	Maryland	5.38	12.40	10.21
	Minnesota	14.05	13.62	4.69
	Missouri	13.45	5.66	4.04
	North Carolina	10.30	18.71	7.89
	Tennessee	7.26	4.18	6.86
	Texas	14.81	2.83	7.29
	Virginia	5.31	7.86	5.07
	Washington	13.25	15.70	13.19
	Average *MAPE*	10.53	9.16	6.94
Labor dominant	Alabama	10.40	7.13	7.38
	Colorado	16.38	10.47	5.77
	Florida	18.94	12.72	3.91
	Indiana	8.61	8.64	10.01
	Kentucky	19.70	9.60	6.08
	Louisiana	7.29	8.51	71.09
	Michigan	15.43	7.91	12.08
	Pennsylvania	14.61	11.20	3.38
	South Carolina	8.87	15.57	16.96
	Wisconsin	17.83	15.28	11.78
	Average *MAPE*	13.81	10.70	8.59*
Overall average MAPE		12.36	9.51	8.49*

*Note: Louisiana was omitted from this calculation

had the smallest *MAPE* (8.49) as compared with the regional membership method ($MAPE = 9.51$) and with the family membership method ($MAPE = 12.36$). From the results reported in Table 5.5 for more populated states, the temporal aggregation method appears to offer the most promise as a method for imposing migration age structure. However, it is clear that each of the methods worked "best" for some states. For 13 of the 26 states, the temporal aggregation method was best. For 9 of the 26 states, the regional membership method gave the best results, and for 4 of the 26 states the family membership method was the winner.

5.4.3 The Imposing Methods Applied to the Less Populated States

We now apply the methods for imposing migration age structure to the ACS 2007 data, targeting the 25 less populated states. Here there is no clear way to choose the best method from the three competing methods. For the less populated states, the ACS 2007 derived model schedules are less reliable than for the more populated states, and, therefore, they cannot be viewed as reflecting the accurate migration age structure of the state. We have developed strategies for deciding which of the imposed schedules is "best" under these circumstances, and these strategies are based on reasoning that when there is correspondence between two schedules it provides evidence of the validity for the method(s) involved. What follows are examples of the logic that underlies these strategies.

Despite a lack of reliability in the ACS 2007 derived model schedules, for the less populated states, they nevertheless can be used to validate one method over another. The three methods were applied to the 25 less populated states, and the three sets of model schedules were contrasted with the model schedules estimated from the ACS 2007 data. Of these states, there were only two migration families identified (child and labor dominant with no retirement peak) and the resulting *MAPE* statistics are reported in Table 5.6. When the *MAPE* is small (less than 10%) it indicates a close correspondence between the imposed model schedule and the ACS 2007 derived model schedule, and this correspondence gives some validity to the imposed model schedule.

For example, there is a close correspondence between the temporally aggregated imposed model schedule and the ACS 2007 derived model schedule for Nebraska ($MAPE = 6.05$) as compared to the family membership method ($MAPE = 13.58$) and the regional membership method ($MAPE = 15.53$). The similarities between the two model schedules are presented in Fig. 5.13, and they suggest there was stability in the migration age structure for Nebraska over the 3-year period from 2005 to 2007. Furthermore, the model schedule imposed by the temporal aggregation method, which is based on the observations for three years of the ACS, inevitably provides a more reliable estimate of the migration age structure than does the ACS 2007 alone.

For Arkansas, there is close alignment between the model schedule imposed by the regional membership method and the ACS 2007 derived model schedule ($MAPE = 8.73$) as compared to the family membership method ($MAPE = 15.16$) and the temporal aggregation method ($MAPE = 9.10$). This is illustrated in Fig. 5.14. It suggests that the migration age structures for Arkansas and the other states in the West South Central Division are very homogeneous. Due to this implied regional homogeneity, we believe that the imposed model schedule, derived from all of the ACS 2007 observations for the Division, provides a more reliable estimate of the migration age structure for Arkansas than the model schedule derived from the ACS 2007 respondents for Arkansas alone.

The state of Iowa is an example of a less populated state with a migration age structure that could be successfully imposed by the family membership method. Iowa was classified as a member of the child dominant (no retirement peak)

Table 5.6 A comparison of the three imposing methods applied to the less populated states: goodness-of-fit statistics (*MAPE*) contrasting the method's model schedule with the model schedule derived from the ACS 2007

	State classifications	Family membership method *MAPE*	Regional membership method *MAPE*	Temporal aggregation method *MAPE*
No retirement peak:				
Child dominant	Alaska	14.44	14.96	8.64
	Arkansas	15.16	8.73	9.10
	Connecticut	21.24	15.55	15.36
	District of Columbia	20.06	14.73	7.72
	Idaho	22.41	34.89	25.71
	Iowa	5.14	11.32	24.67
	Kansas	9.89	11.25	8.36
	Nevada	4.62	12.75	6.69
	New Hampshire	8.00	18.49	9.68
	New Mexico	15.93	11.76	7.77
	Oregon	10.59	13.01	11.88
	South Dakota	27.08	22.65	15.58
	Utah	12.08	13.75	8.14
	Vermont	18.51	17.69	11.43
	West Virginia	10.76	15.63	7.73
	Wyoming	13.09	13.24	7.15
	Average *MAPE*	14.31	15.65	11.60
Labor dominant	Delaware	16.69	22.74	16.36
	Hawaii	13.97	43.77	9.65
	Maine	25.97	30.74	17.57
	Mississippi	11.00	10.87	18.05
	Montana	22.21	28.56	12.15
	Nebraska	13.58	15.53	6.05
	North Dakota	14.68	17.99	18.17
	Oklahoma	16.50	17.10	17.51
	Rhode Island	17.00	18.76	13.86
	Average *MAPE*	16.84	22.90	14.38
Overall average *MAPE*		15.22	18.26	12.60

migration family, and, as reported in Table 5.6, the model schedule imposed by the child dominant family was very similar to the ACS 2007 derived model schedule for Iowa ($MAPE = 5.14$), as compared to the regional membership method ($MAPE = 11.32$) and the temporal aggregation method ($MAPE = 24.67$). This result suggests that Iowa is a conforming member of the child dominant migration (no retirement peak) family, and this lends credence to the argument that the migration age structure imposed by family membership method provides a reliable alternative to the ACS 2007 derived model schedule for Iowa.

What strategy does one use if there is no close correspondence between an imposed model schedule and the ACS 2007 derived model schedule? How can an imposed method be applied? In some of these cases, there is a correspondence between two of the imposed model schedules, and neither is closely aligned with

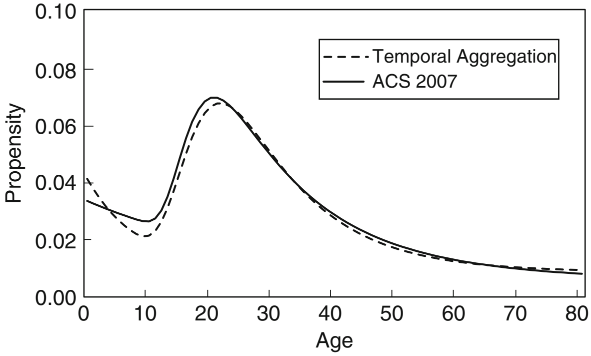


Fig. 5.13 The imposed model schedule for Nebraska demonstrating the correspondence between the temporal aggregation method and the ACS 2007 model schedule

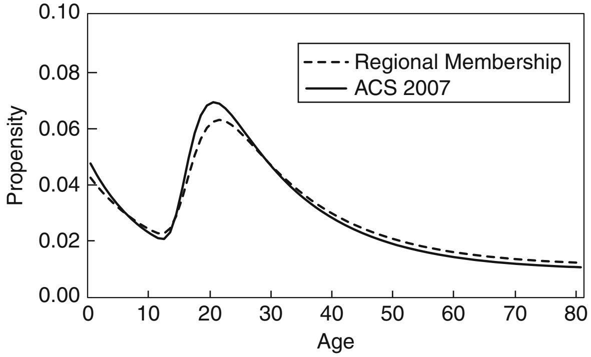


Fig. 5.14 The imposed model schedule for Arkansas demonstrating the correspondence between the regional membership method and the ACS 2007 model schedule

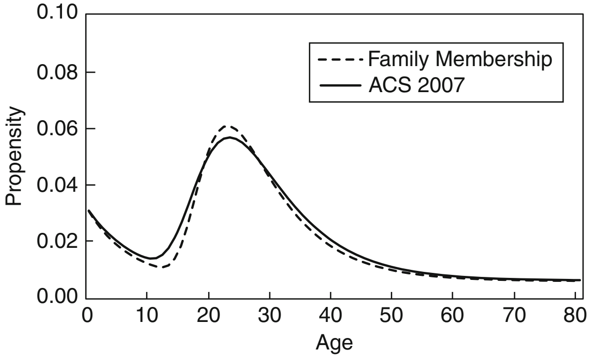


Fig. 5.15 The imposed model schedule for Iowa demonstrating the correspondence between the family membership method and the ACS 2007 model schedule

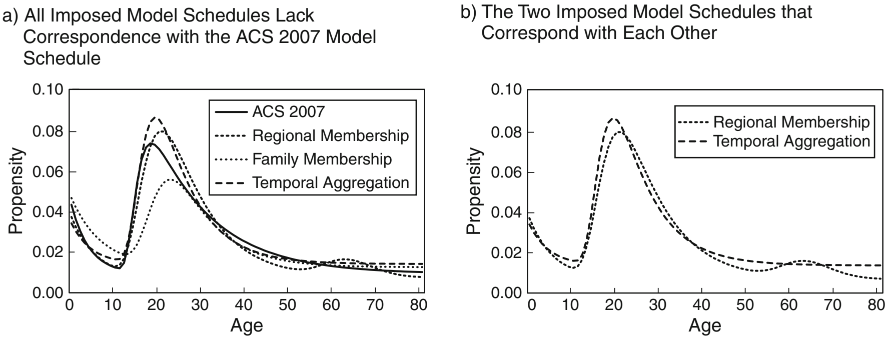


Fig. 5.16 The imposed model schedules for Connecticut lack correspondence with the ACS 2007 model schedule, but two of the imposed model schedules correspond with each other

the ACS 2007 derived model schedule. Connecticut is an example of such a situation. In Fig. 5.16, the top Panel (a) shows there is no close correspondence between any of the three methods and the ACS 2007 model schedule. This is confirmed by the results in Table 5.6 reporting the *MAPE*s for the family membership method (21.24), for the regional membership method (15.55), and for the temporal aggregation method (15.36). In this example, two of the imposing methods have close correspondence, and this is visually apparent in Panel (b) of Fig. 5.16, which shows that the model schedules imposed by regional membership method and the temporal aggregation methods are quite similar. In this case, we advocate using either of these two imposed model schedules as a more reliable alternative to the ACS 2007 derived model schedule. Because of the close alignment between the schedules generated from the regional membership method and the temporal aggregation method, and because of their departure from the ACS 2007 model schedule, we assume that either of these methods offers a more reliable approach for estimating the migration structure than does the direct method based on the ACS 2007 data.

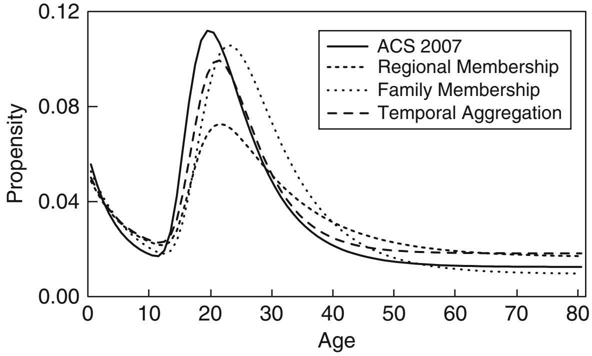


Fig. 5.17 The imposed model schedules for Idaho demonstrating inconclusive results

The results of the experiments, which applied the methods for imposing migration age structures onto the ACS data, were not always conclusive. Idaho, for example, exhibited a migration age structure based on the ACS 2007 data that was dissimilar to other states in its Division, reflected in the regional membership method $MAPE = 34.89$, and the imposed model schedules from the family membership method ($MAPE = 22.41$) and the temporal aggregation method ($MAPE = 25.71$), all of which were equally lacking in correspondence with the ACS 2007 derived model schedule. In this situation there is no clear strategy for imposing the most reliable migration age structure, but the temporal aggregation method seems to be the most consistent and reliable method in general, and, in this case, it provides a model schedule that appears visually, in Fig. 5.17, to be the "average" of all the other model schedules.

5.5 Imposing Spatial Migration Patterns

The imposition of age patterns of migration to "discipline" inadequate data holds great promise for developing improved estimates of in-migration, out-migration, and destination-specific migration flows. In this section, we present a method that adopts a relational perspective, where the age and spatial patterns of migration are related not to a migration family or a standard, but to historical patterns of migration. The historical patterns, and the assumptions regarding trends, are used as a basis for improving observed migration flow data. However, such preliminary "predictions" also could involve a standard. Indeed, the construction of migration flows may involve a combination of information from several data sources (Willekens, 1994). The main feature of this method is that we use a *log-linear model* to capture the contributions of the various data sources. That model provides a convenient way to predict migration from inadequate data, and its parameters define the relative contributions of each of the data sets. The example presented in this section comes from Rogers, Willekens et al. (2003).

5.5.1 Data

We begin by analyzing interregional migration data from two sources: the 1980 and 1990 U.S. Censuses and the 1985 Current Population Survey (CPS). The decennial censuses provide migration data for the periods 1975–1980 and 1985–1990; the CPS provides data for 1980–1985. The data represent numbers of persons by region of residence at time of census or survey and region of residence five years prior to that census or survey. The regions in the analysis are the Northeast, Midwest, South, and West regions, as defined by the U.S. Census Bureau. The 1975–1980 and 1985–1990 migration data are based on a much larger sample size (one of about 1.5 million to 2.0 million persons, i.e., 5% of the U.S. decennial census enumerations) compared to the 1980–1985 migration data (with a sample size of about 50,000 households).

Hence, the accuracy of the latter, understandably, is viewed with some question. Since the adoption of a log-linear model as a vehicle for the indirect estimation of migration relies on an unambiguous interpretation of the model's parameters, the link between the data and the parameters is given particular attention.

5.5.2 Modeling Origin-Destination Migration Flows with Prior Information

The observed migration flow tables and corresponding multiplicative components for the 1975–1980, 1980–1985, and 1985–1990 periods are set out in Table 5.7. The multiplicative components set out in Panel (b) of this table can help us to understand, for example, why the Northeast to South flow in 1980–1985 is so much lower than in the 1975–1980 and 1985–1990 periods. In this case, the reason can be explained by the much lower origin-destination interaction value (i.e., 0.0870 in 1980–1985 versus 0.1144 in 1975–1980 and 0.1120 in 1985–1990). The multiplicative components for the overall level and origin and destination main effects appear reasonable and in line with the census data. In this section, we try to improve the CPS data by imposing the origin-destination association structures found in the census data. When imposing spatial (or age) patterns of migration data, there are two questions researchers should ask before doing so: (1) Can the two (or more) data sets be combined? (2) If so, what structures should be imposed from the auxiliary data?

To illustrate the method of imposing data in a log-linear modeling framework, we predict the 1980–1985 CPS migration flow matrix based on the marginal totals of that data and the spatial structure of the 1975–1980 migration flow matrix implied by the model specified in Eq. (3.4). The resulting predicted migration flow table and ratios of predicted-to-observed migration flow tables for the 1980–1985 period are set out in Table 5.8. Here, we see that the Northeast to South flow increased from 1.4 million to 1.6 million, which is closer to the 1.8 million observed during both census periods. It appears that by imposing historical census data, we have improved the accuracy of the CPS data.

Next, consider the imposition of age patterns. Here, only a dozen 5-year age groups are distinguished in our analysis, ranging from the 0–4 years to 55–59 years. The published CPS data on interregional migration do not provide age detail beyond age 60 (in 1980). Moreover, the published CPS data are for 5-year age groups up to age 34 and for 10-year age groups for ages 35 and higher. To overcome this obstacle, the 10-year age data were disaggregated into 5-year data by assuming a uniform distribution of migrants in the 10-year age interval. The observed age patterns of migration from the Northeast, for example, are set out in Fig. 5.18. Here, we see that the CPS data are incomplete and that they do not always correspond with the patterns found in the censuses.

Given the age-specific patterns observed in the CPS, we assume that none of the age patterns are reliable, and, instead, we borrow these structures, along with the origin-destination associations, from the two censuses. This allows us to basically

Table 5.7. U.S. interregional migration flows (in thousands): 1975–1980, 1980–1985, and 1985–1990 and their corresponding multiplicative components

(a) Flows						
1975–1980	Northeast	43,123	462	1,800	753	46,138
	Midwest	350	51,136	1,845	1,269	54,600
	South	695	1,082	67,095	1,141	70,013
	West	287	677	1,120	37,902	39,986
	Total	44,455	53,357	71,860	41,065	210,737
1980–1985	Northeast	44,845	379	1,387	473	47,084
	Midwest	326	52,311	1,954	1,144	55,735
	South	651	855	68,742	1,024	71,272
	West	237	669	1,085	40,028	42,019
	Total	46,059	54,214	73,168	42,669	216,110
1985–1990	Northeast	44,379	357	1,822	541	47,099
	Midwest	378	52,301	1,766	1,025	55,470
	South	849	1,242	72,887	1,263	76,241
	West	389	705	1,178	43,733	46,005
	Total	45,995	54,605	77,653	46,562	224,815
(b) Multiplicative components						
1975–1980	Northeast	4.4307	0.0395	0.1144	0.0838	0.2189
	Midwest	0.0304	3.6990	0.0991	0.1193	0.2591
	South	0.0471	0.0610	2.8104	0.0836	0.3322
	West	0.0340	0.0669	0.0821	4.8643	0.1897
	Total	0.2110	0.2532	0.3410	0.1949	210,737
1980–1985	Northeast	4.4689	0.0321	0.0870	0.0509	0.2179
	Midwest	0.0274	3.7414	0.1036	0.1040	0.2579
	South	0.0429	0.0478	2.8488	0.0728	0.3298
	West	0.0265	0.0635	0.0763	4.8248	0.1944
	Total	0.2131	0.2509	0.3386	0.1974	216,110
1985–1990	Northeast	4.6055	0.0312	0.1120	0.0555	0.2095
	Midwest	0.0333	3.8819	0.0922	0.0892	0.2467
	South	0.0544	0.0671	2.7678	0.0800	0.3391
	West	0.0413	0.0631	0.0741	4.5898	0.2046
	Total	0.2046	0.2429	0.3454	0.2071	224,815

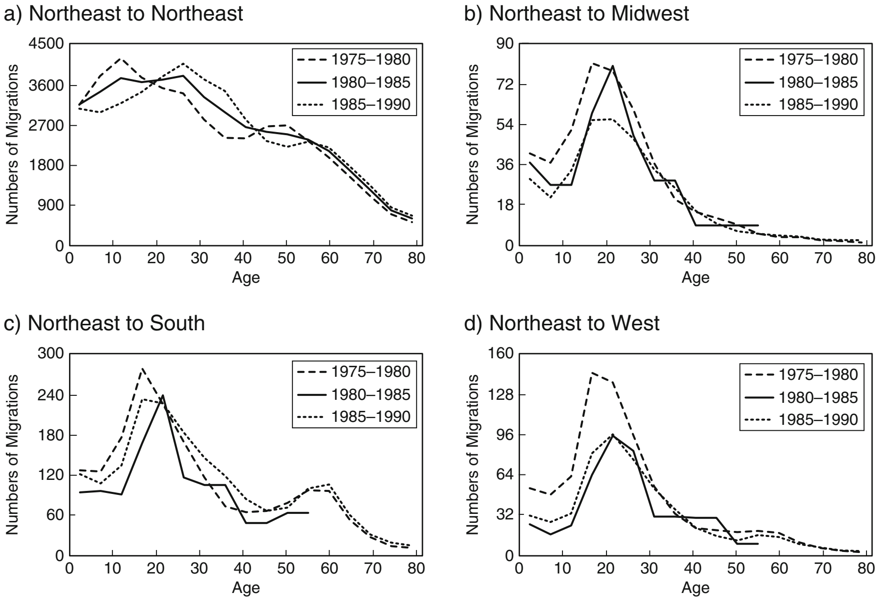
interpolate between the 1980 and 1990 Censuses but with the constraint that the regional origin and destination populations match the CPS. The log-linear-with-offset model for this exercise is specified as:

$$\ln\left(\hat{n}_{ijx}\right) = \lambda + \lambda_i^O + \lambda_j^D + \ln\left(n_{ijx}^*\right), \tag{5.1}$$

where the offset, n_{ijx}^*, represents the interpolated age-specific census flows for the 1980–1985 period. This interpolated data set is forced to fit the origin and destination marginal totals of the 1985 CPS data. The age main effect and interaction structures are borrowed from the interpolated census migration data. The predicted

Table 5.8 The predicted migration flows (in thousands) based on the log-linear model with the 1975–1980 migration flow table as the offset and the ratios of predicted to observed migration flows for the 1980–1985 period

	Region of destination				
Region of Origin	Northeast	Midwest	South	West	Total
(a) Predicted flows					
Northeast	44,445	393	1,614	632	47,084
Midwest	431	52,055	1,977	1,272	55,735
South	814	1,047	68,324	1,087	71,272
West	369	719	1,253	39,678	42,019
Total	46,059	54,214	73,168	42,669	216,110
(b) Ratios of predicted to observed					
Northeast	0.9911	1.0369	1.1637	1.3362	1.0000
Midwest	1.3221	0.9951	1.0118	1.1119	1.0001
South	1.2504	1.2246	0.9939	1.0615	1.0000
West	1.5570	1.0747	1.1548	0.9913	1.0000
Total	1.0000	1.0001	1.0000	1.0000	1.0000



a) Northeast to Northeast

b) Northeast to Midwest

c) Northeast to South

d) Northeast to West

Fig. 5.18 Observed age-specific migrations from the Northeast: 1975–1980 (1980 Census), 1980–1985 (Current Population Survey), and 1985–1990 (1990 Census)

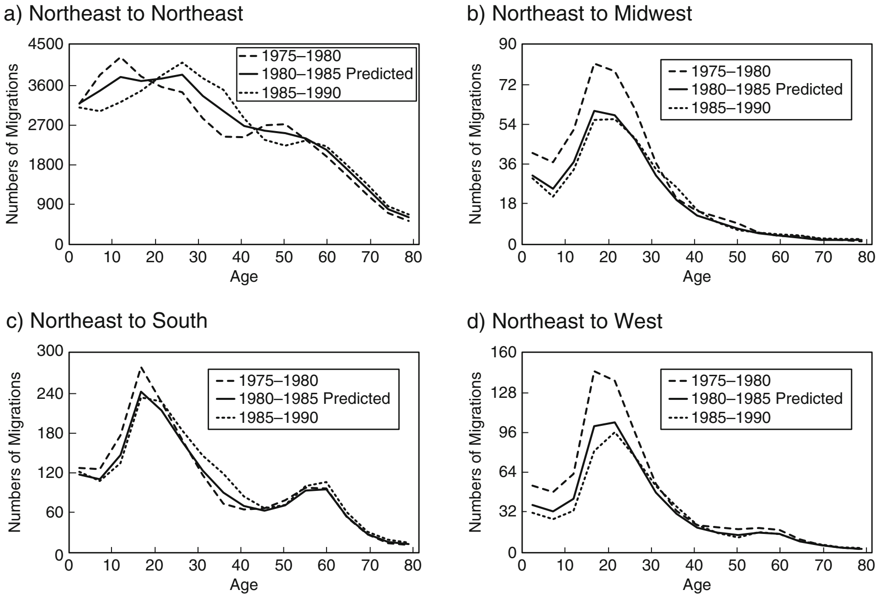


Fig. 5.19 Age-specific migrations from the Northeast: Observed 1975–1980, predicted 1980–1985 (using interpolated 1975–1980 and 1985–1990 flows as the offset in a log-linear model), and observed 1985–1990

flows are set out in Fig. 5.19 for migration from the Northeast. It is evident that, in this case, the predicted patterns correspond closely with the 1980 and 1990 Census migration patterns, whereas the observed CPS data do not.

5.6 Summary and Discussion

This chapter has focused on methods for "repairing" inadequate migration data, particularly data obtained from smaller sample sizes that are insufficient to yield reliable age-specific migration flows. Our methods are similar to the three most popular procedures used to deal with inadequate mortality data. First, we can borrow more reliable schedules from larger areal units that include the geographic area under consideration (e.g., counties within states, or states within divisions). Second, we instead can borrow one of a set of "standard" schedules that represent different "families" of schedules (e.g., families with or without a retirement peak). Third, we can simply aggregate the available data for the geographic area under analysis over several years. We have called these three methods the regional membership method, the family membership method, and the temporal aggregation method.

As in Chapter 4, our methods are first tested, where possible, using Census 2000 PUMS 1% sample data, with results that are compared to those obtained from the Census 2000 full sample data. And we have applied our methods for repairing data to the data collected by ACS. Our conclusion is that each of the three methods tends to work best in different situations.

The indirect estimation of the levels and age patterns of fertility and mortality has a long history in demography. A dominant strategy there has been to combine empirical regularities with other information to fill-in the missing data. Functional representations (Heligman & Pollard, 1980) and relational representations (Brass, 1974) of age patterns have occupied a central position in such efforts at indirect estimation (Preston, Heuveline, & Guillot, 2001). The indirect estimation of migration is of a more recent date, in part because the problem is more complicated. For example, the age patterns of migrants depend on the direction of migration. To be acceptable, therefore, a method must somehow integrate the age pattern with the spatial pattern. Section 5.5 proposes such a method, and there we outline a very general log-linear model for imposing structure on inadequate observed migration flow data.

Our general approach has been one that uses a model to impose migration structure from partial data contributed by different data sources. The various explanatory variables that are commonly used in such models are replaced by different data sources. When the problem is to predict the *number* of migrants by origin, destination, and age, the appropriate model is the log-linear model. The log-linear model becomes a vehicle to determine if the distribution of counts among the cells of a table can be accounted for by an underlying structure. If the data are incomplete, the underlying structure is determined by data availability, with the parameters of the log-linear model identifying the contributions of the various partial data sets to the predicted migration flows.

Chapter 6
Inferring Age and Spatial Patterns

6.1 Introduction

In this chapter, we focus on methods for estimating migration flows in the absence of migration data. To obtain the patterns of interest, we use auxiliary information. Our examples illustrate both current and historical applications of indirect estimation. In Section 6.2, a model for estimating the age composition of out-migration in the United States from aggregate totals of out-migration and population age compositions is presented. This work draws from a recent paper by Little and Rogers (2007). The possibility of using 0–4 year old birthplace-specific population stocks to estimate interregional migration flows is demonstrated in Section 6.3, following work set out in Rogers and Jordan (2004) and Raymer and Rogers (2007). We then apply the methodology to estimate the historical (and completely missing) migration flows for the 1905–1910 and 1915–1920 periods. Finally, in Section 6.4, we focus on the potential for merging migration data obtained from multiple sources. Here, the aim is to follow Frans Willekens's recommendation that "in order to compile coherent and internally consistent information on migration, data from several sources ought to be combined" (Willekens, 1994, p. 31). Smith, Raymer, and Giulietti (2010), for example, follow this advice by combining census, registration, and survey migration data in England and Wales. For *our* illustration, in Section 6.4 we combine migration data obtained from the Internal Revenue Service (IRS) and the American Community Survey (ACS) PUMS samples. Both sources provide annual information on migration, but with different levels of measurement and attributes. We use the IRS data to improve the spatial patterns of the ACS PUMS data. The result is a synthetic database that exhibits more stable migration patterns over time.

6.2 Age Compositions of Out-Migrants

In this section, we argue that the age distribution of a population provides valuable information about the age composition of its out-migrants, and we propose a method for estimating the age profile of out-migrants when there may be insufficient data. At the outset, it should be noted that we leave the task of estimating the number

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_6,

of out-migrants to others, and for the purposes of this section we assume accurate information about total out-migration flows is available. Also, although we recognize that past in-migration flows influence a population's current age structure as well as the age structure of its in-migration streams, we do not, at this time, attempt to introduce the possible impacts of such flows. Indeed our objective is a modest one, namely, to offer a method that uses the characteristics of the age distribution of a population to predict the most likely age profile of its out-migrating population.

In pursuit of this goal, we have three preliminary objectives. The first of these is to develop the thesis that the age distribution of a population will inevitably influence the age profile of the out-migrating population. The second objective addresses questions about the regularities of the age profiles of out-migration. Here, we demonstrate that profiles of the age composition of migrants often can be simplified without loss of information by adopting the Rogers-Castro model migration schedule. We show that the 7-parameter model migration schedule (see Chapter 2, Section 2.2.2) adequately represents out-migration profiles for different geographic scales, including states, consolidated metropolitan areas (CMSAs), metropolitan areas (MSAs), and non-metropolitan counties, and that the variation in age profiles across geographic units can be captured by a simple typology of model schedules. The final step is to show how the characteristics of the age distribution of a population can be used to predict the most likely type of model schedule for the out-migrating population.

6.2.1 Data

The age compositions of migrants were obtained for four different U.S. geographic areas, all generally large-scale, but decreasing in population size: 51 states (including the District of Columbia), 18 CMSAs, 258 MSAs, and 3,101 non-metropolitan counties. The age-specific out-migration data for these geographic areas come from the Census 2000 Migration DVD provided by the US Census Bureau. It gives counts of persons who left their area of residence between 1995 and 2000 and lived to be counted as residing in another area by the 2000 Census. Based on a person's age in 2000, these counts are disaggregated into 5-year age categories, beginning at age 5 and ending at age 85 or older. The age composition of the out-migrating population (i.e., $N_i(x)$, where i denotes the place of origin and x denotes the age at the beginning of the age interval) is determined by the number of migrants in each age category divided by the total number of migrants.

There are 17 age categories for each out-migration schedule, representing profiles that are not smooth and, indeed, are fairly coarse. So the 17 age groups were converted from 5-year age intervals to single-year age intervals by dividing by five and assigning this value to the middle single year age group. For example, for the 5-year age interval 15–19 years, the value would be assigned to the 17 year-old age group. Cubic spline interpolations were then used to arrive at smooth profiles for all integer values of x, using the middle single-year age groups as the nodes for the spline algorithm, carried out by the Advanced Systems and Design add-on function in Microsoft Excel. The entire process transformed the observed 5-year age compositions of migrants into the corresponding single-year age compositions.

The population data came from the U.S. Census Bureau's intercensal population estimates for 1995. These are provided for all states and counties with population breakdowns for age categories, 0, 1–4, 5–9,. . .85+. The first two age groups were combined into a 5-year category 0–4 to give a total of seventeen 5-year age categories, 0–4, 5–9, 10–15, . . .80–84, 85+. The county populations were then aggregated to form the MSA and CMSA counts according to the U.S. Office of Management and Budget (OMB) definitions of 1999. Non-metropolitan counties are those counties that are not part of an MSA or a CMSA. Note, counties with fewer than 100 persons in any one of the 5-year age categories were excluded. This was based on an arbitrary decision, which deemed that if there were fewer than 100 people at risk for migration the out-migration profiles would be too unstable for consideration. This strategy left 1,944 non-metropolitan counties in the study, and a total of 2,271 out-migration schedules across the four geographies.

6.2.2 Relationship Between Population Age Structures and Migration Age Structures

Previous work has applied the Rogers-Castro model migration schedule to national and regional migration profiles, and to total flows and well as directional flows. Here we extend this work by testing, in a rigorous way, if the seven-parameter model is generally effective for representing the age composition of migrants, and at what geographic scales. Note, the 9-, 11- and 13-parameter extensions to the 7-parameter model schedule are more common for migration rate schedules and for schedules that represent directional flows. For the age compositions of total (non-directional) out-migration, the post-labor force population generally is not a large enough proportion of the population to warrant the additional complexity of the 9-, 11- or 13-parameter model schedules.

The fitting to the 2,271 single-year age compositions (across four geographies) was carried out with a customized SPSS program that estimated the seven parameters of the Rogers-Castro model schedule. After fitting the model schedule to the age composition profiles that derived from the cubic spline interpolation of the observed data, the R^2 was used to evaluate if the splined profiles conformed to the model schedule. Values of R^2 equal to 0.94 or greater were considered good fits. All of the states and CMSAs, and 254 of the 258 of the MSAs, met this baseline goodness-of-fit criterion.

Of the initial 1,944 non-metropolitan counties that had more than 100 people in each of the 5-year age categories, only 987, or 51%, were satisfactorily fitted by the Rogers-Castro model. This was a clear indication that the initial filter that excluded low population counties was not sufficiently large enough, and that there remained irregular migration age profiles from counties with few people at risk for migration. Little and Rogers (2007) found that as population size increases so does the percentage of counties with out-migration schedules that conform to the

Rogers-Castro model. Of the counties with fewer than 30,000 people, only 38.5% (517 of 1,342) had out-migration profiles that were adequately fitted by the model schedule, but among the counties with more than 30,000 people, 78.2% (471 of 602) had out-migration profiles that conformed to the model schedule. For that reason it was decided that a minimum population size of 30,000 might be a reasonable expectation for this application of the model migration schedule.

A goodness-of-fit summary for the *MAPE* statistics as well as the R^2 values is reported in Table 6.1. In this particular application, which fits model schedules to age-specific proportions of the total out-migrating population, the *MAPE* statistic can overstate the error between the observed and the fitted schedules. For example, Fig. 6.1 displays the MSA (Bryan-College Station, Texas) and the county (Montgomery County, Virginia) with *MAPE* values that were the largest found among those with conforming schedules. The large *MAPE* values of 31.16 for Bryan-College Station and 31.31 for Montgomery County, Virginia are not consistent with the degree of correspondence that is indicated by the R^2 statistics and

Table 6.1 Distributions of the R^2 and *MAPE* statistics for the areas with model migration schedules that conformed to the model migration schedule, by geographic scale

	States $N=51$	CMSAs $N=18$	MSAs $N=254$	Counties $N=471$
R^2				
Mean	0.99	0.99	0.98	0.97
Min	0.98	0.97	0.94	0.94
Max	1.00	1.00	1.00	1.00
SD	0.01	0.01	0.01	0.01
MAPE				
Mean	6.79	8.64	11.91	15.57
Min	2.74	4.86	2.25	6.09
Max	14.30	16.08	31.16	31.31
SD	2.74	3.05	5.73	5.08

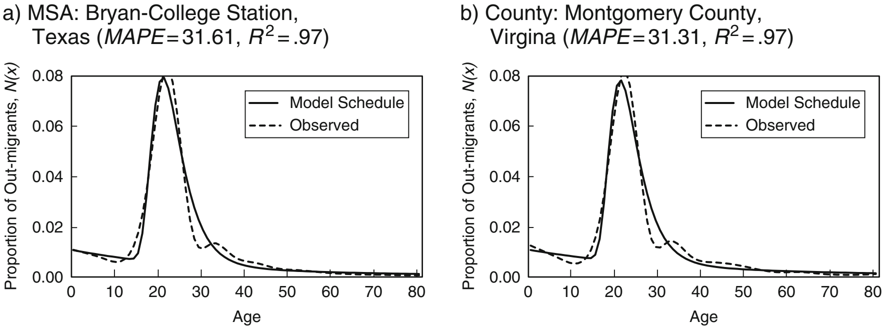


Fig. 6.1 Examples of out-migration schedules that conform to the model schedule but have large *MAPE* values

from what is visually apparent in Panels (a) and (b) of Fig. 6.1. In both cases, the out-migrating populations are dominated by the college-aged populations, and, in contrast, the proportions of out-migrants in the oldest ages are close to zero. When a denominator is near zero it causes numerical instability, and even a small difference between the observed and the fitted values will result in a large contribution to the *MAPE*. For this reason, the R^2 statistic was chosen as a better indicator of conformity to the model schedule than the *MAPE*.

Figure 6.2 reveals more about what it means for age composition profiles to conform to the model schedule. Sarasota-Bradenton, Florida, is the MSA that had the age composition of the out-migrants that barely satisfied the cut-off criteria for conformity, and Franklin, Illinois, is the conforming county that came the closest to being classified as nonconforming. If an MSA or a county had observed and fitted age profiles that were more disparate than those displayed in Fig. 6.2, it was treated as nonconforming.

From the proportion of units within each geographic set that were classified as conforming to the Rogers-Castro model schedule, and from the goodness-of-fit statistics presented in Table 6.1, it is clear that the larger-scale geographic units tend to have observed out-migration schedules that conform more closely to the estimated model schedule. On average, those states have the highest degree of fit measured by their high mean R^2 values and *MAPE* statistics and the corresponding low standard deviations. The conforming counties have the lowest R^2 values and *MAPE* statistics, on average, and the largest deviations, suggesting that, in general, smoother and simpler observed age profiles are more likely to be found in geographic units with larger populations.

The four MSAs with observed profiles that did not produce satisfactory fits by the model schedule were Punta Gorda, Florida ($R^2 = 0.65$), Fort Pierce-Port St. Lucie, Florida ($R^2 = 0.90$), Odessa-Midland, Texas ($R^2 <= 0.92$), and Victoria, Texas ($R^2 = 0.93$). These MSAs have distinctly unusual population compositions that are dominated by retirement communities or military bases, which gives further

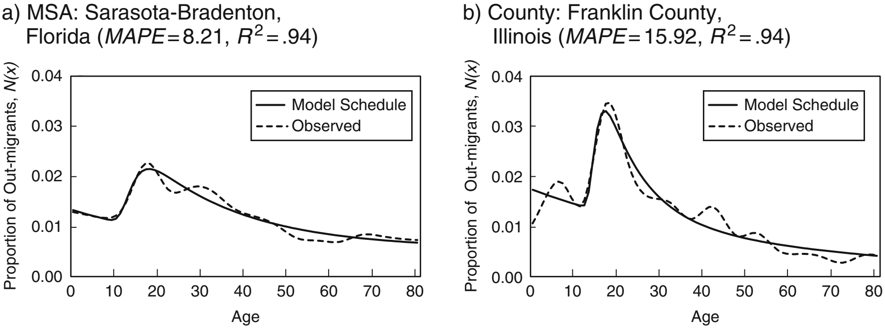


Fig. 6.2 The MSA and county out-migration schedules that meet the minimum standard for conforming to the model schedule

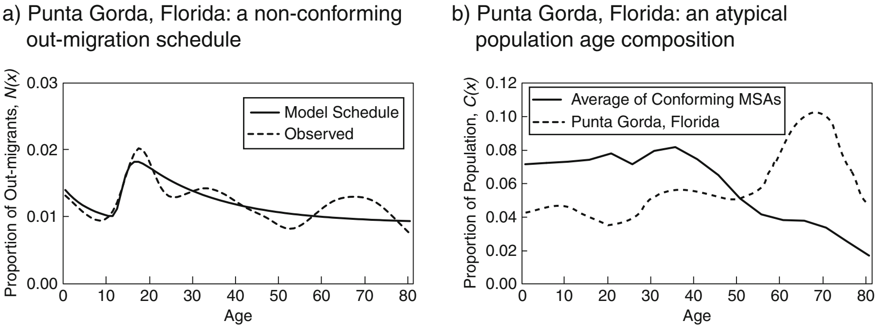


Fig. 6.3 MSA Punta Gorda, Florida: A demonstration of an atypical population age composition associated with a non-conforming out-migration schedule. (*Source*: Little & Rogers, 2007, Figs. 6.4 and 6.5)

credence to the thesis that the out-migration profile is determined, in part, by the age composition of the population of origin. The Punta Gorda, Florida, MSA was chosen to demonstrate a nonconforming out-migration schedule as displayed in Panel (a) of Fig. 6.3. That model migration schedule clearly does not capture the heavy representation of older aged migrants among the total population of out-migrants. Panel (b) shows that the population composition of the non-conforming Punta Gorda, Florida, MSA is disproportionately older when compared to the average of the population compositions of the 254 conforming MSAs. These findings also suggest that the Rogers-Castro model migration schedule may be appropriate only when a population distribution shows decreasing proportions in the older ages, or, at least, when it conforms to what might be called a typical population age composition.

In summary, the model migration schedule seems to be a very effective tool for representing out-migration schedules when population sizes are large (i.e., above 30,000) and when the age distribution of a population is not particularly unusual, such as those that represent a military base or a large retirement community.

6.2.3 Typologies of Model Migration Schedules

To reduce the complexity of the variation among model migration schedules, we took the parameters associated with 794 model schedules (now 51 states plus 18 CMSAs, 254 MSAs, and 471 non-metropolitan counties with populations greater than 30,000) and clustered them according to their parameter values, using the k-means method in the QUICK CLUSTER procedure in SPSS (2004). Initially, two clusters are formed and their centers are defined by the sets of parameters that are farthest apart (in Euclidean distance), and the rest of the areas are assigned to the cluster that is nearer. The new cluster centers are calculated as the mean parameter values of all areas assigned to that cluster. All cases are assigned to the new cluster centers again and the process is repeated. The Bayesian Information Criterion (Raftery, 1995) was used to determine the optimal number of clusters for each of the geographies.

The 51 states and 18 CMSAs were combined, because of their similarity in population size and because of the small number of CMSAs, and the set of 69 schedules was reduced to two types of schedules—the "Standard" model schedule and the "Delayed Career" model schedule. These are presented in Panel (a) of Fig. 6.4. The Standard schedule represents the migrant age composition for the majority of the states and CMSAs ($N = 59$). The remaining schedules ($N = 10$) fell into the Delayed Career cluster, with a career migration peak that is flatter and peaking a few years later than in the Standard model schedule. The parameter values associated with the two model types are reported in Table 6.2, and the standard deviations quantify the variation of each parameter within each cluster. The standard deviations are generally small and suggest that the individual schedules that fall within a cluster are similar to the schedules defined by the cluster centroids. The R^2 values representing the goodness-of-fit between the model schedule that defines a cluster and the model schedules for the states and CMSAs within that cluster were calculated, and the Standard cluster which contains most states and CMSAs had an average

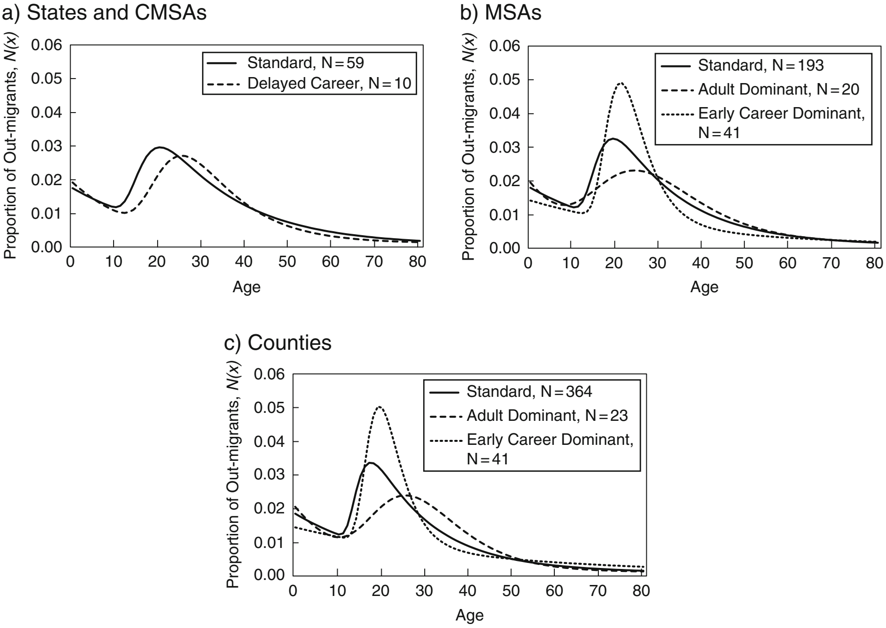


Fig. 6.4 Typologies of model out-migration schedules. (*Source*: Little & Rogers, 2007, Fig. 6.6)

Table 6.2 Clusters of model migration schedules and the parameter values that define the centroids of the clusters by geographic scale

(a) States and CMSAs		a_1	α_1	a_2	α_2	μ_2	λ_2	c
Standard ($N = 59$)	*Mean*	0.0171	0.0439	0.0377	0.0590	15.4014	0.2939	0.0002
	SD	0.0023	0.0151	0.0071	0.0131	1.8478	0.1119	0.0004
Delayed career ($N = 10$)	*Mean*	0.0182	0.0653	0.0557	0.0921	22.3890	0.1587	0.0008
	SD	0.0035	0.0326	0.0069	0.0192	4.1079	0.0940	0.0013
(b) MSAs		a_1	α_1	a_2	α_2	μ_2	λ_2	c
Standard ($N = 193$)	*Mean*	0.0174	0.0423	0.0409	0.0707	14.9789	0.3575	0.0003
	SD	0.0035	0.0177	0.0115	0.0239	2.0901	0.1503	0.0010
Early career dominant ($N = 41$)	*Mean*	0.0138	0.0278	0.1088	0.2061	19.8660	0.2691	0.0001
	SD	0.0031	0.0182	0.0308	0.0419	1.5498	0.0513	0.0005
Adult dominant ($N = 20$)	*Mean*	0.0182	0.0923	0.0541	0.0900	25.7573	0.0854	0.0011
	SD	0.0041	0.0356	0.0085	0.0136	3.9129	0.0262	0.0013
(c) Counties		a_1	α_1	a_2	α_2	μ_2	λ_2	c
Standard ($N = 364$)	*Mean*	0.0178	0.0436	0.0397	0.0803	13.9634	0.4714	0.0005
	SD	0.0039	0.0277	0.0124	0.0342	1.4227	0.1720	0.0010
Early career dominant ($N = 84$)	*Mean*	0.0141	0.0226	0.1095	0.2347	18.6147	0.2958	0.0002
	SD	0.0028	0.0081	0.0316	0.0251	2.2705	0.0644	0.0006
Adult dominant ($N = 23$)	*Mean*	0.0195	0.0810	0.0545	0.1131	28.2344	0.0938	0.0009
	SD	0.0072	0.0415	0.0121	0.0239	2.9721	0.0397	0.0010

Source: Little and Rogers (2007, Tables 6.3, 6.5 and 6.7)

$R^2 = 0.96$, which indicates a very tight cluster. The Delayed Career cluster is less tight with an average $R^2 = 0.91$, which is consistent with the larger standard deviations associated with the parameter values of the Delayed Career cluster reported in Table 6.2.

The same clustering technique produced a three cluster typology for the MSAs. Once again there was a Standard cluster that included most MSAs ($N = 193$). The smaller clusters were the "Early Career Dominant" ($N = 41$) and the "Adult Dominant" ($N = 20$) clusters. The three cluster profiles are presented in Panel (b) of Fig. 6.4, and the parameter values and the variations in parameters within each cluster are reported in Table 6.2. Most of the visible differences between clusters occur around the career migration years. The peak proportion of out-migrants is youngest, around age 18, for the Standard cluster, as compared to age 20 for the Early Career Dominant cluster and age 28 for the Adult Dominant cluster. The pre-labor slope is steepest for the Early Career Dominant cluster and almost as steep for the Standard cluster, whereas the pre-labor slope is clearly flatter for the Adult Dominant cluster. The Early Career Dominant cluster consists of age profiles with high proportions of young adults, who migrate without children, and supporting evidence comes from its relatively low infant migration peak. The Adult Dominant cluster has the flattest and lowest profile during the career migration years, as well as the highest infant migration peak. Together these imply that relatively more adults are migrating at older ages and with children.

In the Standard cluster $\lambda_2 = 0.36$, and its standard deviation, $SD = 0.15$, is the largest relative to the size of the parameter, suggesting that there is variation in the individual schedules in the slope ascending to the labor peak. The Early Career Dominant cluster has the most variation in the a_2 and α_2 parameters, suggesting that there is some variation in the peakedness of the schedules and in the rates of decrease in the older adult ages. The Adult Dominant cluster has the most variation in the pre-labor slope (α_1) parameter and in the constant level parameter (c). When the model schedules for each of the MSAs within the cluster were compared with the model schedule that defines the cluster, the Standard cluster had the least variation with an average $R^2 = 0.95$ and the smallest standard deviation ($SD = 0.04$). The Adult Dominant cluster was also fairly tight (average $R^2 = 0.94$ and $SD = 0.05$). The Early Career Dominant cluster had the most variation (average $R^2 = 0.66$ and $SD = 0.10$).

The out-migration typology for the non-metropolitan counties (displayed in Panel (c) of Fig. 6.4) can be described in a way that is very similar to the typology developed for the MSAs. The same profiles appear: Standard ($N = 364$), Adult Dominant ($N = 23$), and Early Career Dominant ($N = 84$). The counties, like the MSAs, and the state and CMSAs combined, have a Standard cluster with the best average goodness-of-fit (average $R^2 = 0.93$) with the lowest standard deviation ($SD = 0.03$), and like the MSAs, the Early Career Dominant cluster has the most variation within (average $R^2 = 0.83$) and the largest standard deviation ($SD = 0.11$).

6.2.4 Prediction of Migration Family Membership from Population Data

Figure 6.4 reveals that the most visible differences in the families of model schedules occur during the career migration years, especially with regard to the migration of the young and middle-aged adults. Little and Rogers (2007) constructed population measures to account for the variations in the career migration patterns revealed by the clusters in Fig. 6.4. These were generally simple measures of the proportion of the population in the years 20–24, 25–29, and 30–34, when adult migration has the highest propensity. In addition, they developed other population measures designed to affect the labor force slope and migration levels in the early and adult years as well as in infancy and at the oldest ages.

Using these population variables, Little and Rogers (2007) estimated canonical discriminant functions that weighted the population measures to optimally discriminate between the model schedule clusters. Separate models were estimated for the states and the CMSAs combined, the MSAs, and the non-metropolitan counties. For all three levels of geography, the discriminant function analyses were quite successful in predicting the correct migration cluster from the hypothesized population variables. They report that 94.2% of the states and CMSAs, 88.2% of the MSAs, and 88.3% of the counties were successfully classified into one of the three clusters, Standard, Adult Dominant, and Early Career.

A discriminant function analysis was carried out for each of the three levels of geography. For the MSAs and the counties, two discriminant functions were needed to predict membership in the three clusters. Overall, the discriminant analysis method was found to be accurate, but quite complicated. The complete explanation of the results can be found in Little and Rogers (2007). Here, we present a simpler alternative that uses population data to predict membership in the Standard cluster. Figure 6.3 clearly shows that the Standard migration schedule is the most common classification, and the profile of the Standard schedule is similar across all three levels of geography. In view of that, we have designed more practical tools that use data on a population's age structure to help make decisions about the appropriateness of predicting a Standard versus a Non-standard migration schedule. In addition, these methods are designed to be applied uniformly to states, CMSAs, MSAs, and counties.

The first method does not use any information about the age composition of the population and simply assigns the Standard migration schedule to every area. It will be referred to as the "baseline method." Since the Standard schedule is the most common, all 794 of the areas could be classified as Standard, and 59 states and CMSAs, 193 MSAs, and 364 counties, for a total of 616, or 77.48% of all 794 areas, would be categorized correctly in the Standard cluster. That leaves 179 of 794 (22.58%) misclassified. If all states and CMSAs were classified as Standard, this would place all 69 in the Standard cluster, and 59 of the 69 (85.51%) would be correctly classified, and only 10 of the 69 (14.49%) would be incorrectly classified. Similarly, all MSAs and counties could be classified as having a Standard profile, and this would be successful 75.98 % of the time for the MSAs (193 of 254) and 77.28 % of the time for the counties (364 of 471).

The simplicity of the baseline method is attractive, and assigning the Standard migration schedule to every area is a reasonable approach, especially for the states and CMSAs where the baseline method yields success 85.51% of the time. (The effectiveness of the baseline method is summarized in Table 6.3.) However, we offer an alternative that makes use of population data and improves the predictions without adding much complexity. This method is based on the logistic regression model, which was estimated first with all states, CMSAs, MSAs, and counties included, and all of the population variables developed by Little and Rogers (2007) were used as predictors of the binary outcome Y, where $Y = 0$ if the area is a member of the Non-standard cluster, and $Y = 1$ if the area is a member of the Standard cluster. Only three population variables were statistically significant, X_1 = the proportion of the population aged 20–24, X_2 = the proportion aged 25–29, and X_3 = the proportion aged 30–34. The estimated coefficients are $\beta_0 = 5.00$, $\beta_1 = -78.85$, $\beta_2 = 123.18$, $\beta_3 = -80.44$. If the logistic regression model is transformed into its probability form as specified in Eq. (6.1), the "logistic regression method" applies Eq. (6.1) by substituting the proportions of the population aged 20–24, 25–29, and 30–34 for X_1, X_2, and X_3 and the given estimates for β_0, β_1, β_2, and β_3:

$$P(Y = 1) = \frac{1}{(1 + \exp -(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3))} \tag{6.1}$$

A calculation of Eq. (6.1) yields the probability that the Standard migration schedule is the appropriate type for a specific geographic area. If the probability is greater than or equal to 0.50, the area is then predicted to have a Standard migration schedule. When this method was applied to all of the areas, 86.92% (690 of the 794 areas) were classified successfully into Standard and Non-standard clusters.

Separation into the three levels of geography generated three different sets of estimated coefficients. For the 69 states and CMSAs, the logistic regression model

Table 6.3 A comparison of the methods for predicting membership in the Standard and Non-standard out-migration clusters

	Baseline method	Logistic regression method	
	Percent correct	Percent correct	Percent improvement
All geographies ($N = 794$)	77.48	86.92	9.43
States and CMSAs ($N = 69$)	85.51	92.75	7.25
MSAs ($N = 254$)	75.98	86.61	10.63
Counties ($N = 471$)	77.28	87.47	10.19

coefficient estimates are $\beta_0 = 31.06$, $\beta_1 = -301.01$, $\beta_2 = 155.44$, $\beta_3 = -240.35$, and application of the logistic regression method was successful in classifying 92.75% (64 of 69) of the states and CMSAs into the correct Standard and Nonstandard clusters. This contrasts with 85.51% (59 of 69) successful classifications from the baseline method, which demonstrates that the added population data improved the predictions by 7.25%.

The logistic regression coefficient estimates for the MSAs are $\beta_0 = 7.82$, $\beta_1 = -91.73$, $\beta_2 = 144.31$, $\beta_3 = -121.64$, and, when the logistic regression method was applied to the 254 MSAs 86.61% (220 of 254) were correctly classified—a 10.63% improvement over the baseline method. The logistic regression coefficients for the county model are $\beta_0 = 2.66$, $\beta_1 = -76.96$, $\beta_2 = 136.31$, $\beta_3 = -62.45$, and the county results of the logistic regression method were very similar to the MSA results. Indeed, 87.47% (412 of 471) were classified successfully, and this was an improvement of 10.19% over the baseline method. The success of the logistic regression method is summarized and contrasted with the baseline method in Table 6.3.

In conclusion, the discriminant function analyses and the methods outlined in Little and Rogers (2007) offer the most complete and accurate procedures for making use of population data to predict the form of the model schedule that best represents the migrating population. However, we demonstrate that the age composition of out-migrants can be inferred effectively from applying the logistic regression method, which requires only the proportions of the population at risk for migrating aged 20–24, 25–29, and 30–34 along with the logistic regression coefficients provided here.

6.2.5 Summary

Our investigation of the age composition of migrants, and how it is related to the age composition of the population that they left, has led to some fruitful prospects for a method of *indirectly* estimating the age composition of out-migrants. The use of the Rogers-Castro model schedule is central to our method, and, from the findings set out in Section 6.2.2, it seems that most geographic units have out-migration age profiles that are very well represented by that model schedule. The probabilities of identifying out-migration age profiles that conform to it are greatest for the largest geographic units (states and CMSAs). They are still high for MSAs, but the probabilities decrease for non-metropolitan counties, especially those with populations below 30,000. In geographic units where there is a relatively large population and an observed age profile of out-migration that does not conform to the model schedule, the lack of conformability often can be anticipated by a simple inspection of the age composition of the population. For example, if the population distribution has a dramatic bimodal shape in the years of career migration (Little & Rogers, 2007), or, if the proportions of the population in the older age categories increase with age, the probabilities of out-migration age profiles conforming to the Rogers-Castro model migration schedule are diminished.

For those geographic areas with out-migration age profiles that are well represented by the Rogers-Castro model schedule, the finding that most of the variation across the age profiles can be captured with a parsimonious set of profiles is very significant. Knowing that most out-migration age profiles exhibit the Standard shape, which can be summarized by the 7-parameter model schedule, suggests that the age compositions can be estimated indirectly for most of the geographic areas examined here. The proposed estimation procedure comes in two parts. As a first step, reject any geographic area that has an abnormal population age composition as described above, and reject any area that has a population size less than 30,000 people. As a second step, measure the important population composition variables and follow the procedures set out by Little and Rogers (2007) for the most accurate method of assigning a geographic area to the correct cluster, or, alternatively, use the simplified logistic regression method for classifying areas as having a Standard or a Non-standard migration schedule. Once the most likely cluster membership is established, evaluate the model schedule for each age, using the parameter values that are the centroids of that cluster, as set out in Table 6.2.

In conclusion, this section sets out four important findings. First, an examination of over 2,000 age compositions of migrant outflows reveals that a significant majority of them conform to the shape of the reduced 7-parameter form of the Rogers-Castro model migration schedule. Second, geographic scale apparently does matter: large regions with more than 30,000 people (e.g., states, CMSAs, and MSAs) conformed most frequently, whereas smaller area units, such as counties with fewer than 30,000 people, conformed the least number of times. Third, a typology of conforming model schedules is identified that describes the age compositions of a large number of the conforming schedules, at all levels of geography. Finally, it is demonstrated that one can predict with considerable success, the "family" membership of a non-observed age composition of migrants by examining the age composition of the origin population, thereby providing evidence that the two age compositions are linked. The age composition of a population can indeed tell us something about the age composition of its out-migrants.

6.3 Inferring Historical Spatial Patterns Using Infant Migration Estimates

Historical censuses in the United States prior to 1940 did not collect 5-year migration data. Age-specific data by place of residence by place of birth, however, were collected for some of the earlier censuses at the beginning of the 20th century. In this section, we show how these types of data can be used to infer 5-year migration data, following the work by Rogers and Jordan (2004) and Raymer and Rogers (2007), who demonstrated with data obtained from the 1990 and 2000 censuses in the United States and Mexico that reported 0–4 year old birthplace-specific population stocks can be used to estimate 5-year migration patterns of all age groups. After describing some of this work, we then demonstrate how this method can be used to infer historical patterns of internal migration. In particular, our application uses data on U.S.-borns obtained from the 1900, 1910, and 1920 censuses to estimate the 1905–1910 and 1915–1920 migration flows between the Northeast, Midwest, South, and West regions of the U.S.

6.3.1 Estimates Based on Infant Migration Data: The Regression Method

The first set of model mortality schedules published by the United Nations summarized the age-specific death rates of 158 life tables of national populations by using

> . . .regression equations which linked the probability of death in each five-year age interval with the corresponding probability in the previous age interval. . .. Thus model schedules could be calculated by assigning alternative probabilities of infant death from very high to very low, and associating with each . . . the schedule of death probabilities in successive groups calculated from the corresponding regressions. (Coale & Trussell, 1996, p. 475)

The set of life tables so developed would be appropriate for describing the mortality schedule of a particular population as long as the age patterns of death rates were similar in different populations at roughly the same level of mortality, and so long as the 158 life tables were based on reasonably accurate data.

We have carried out exploratory efforts to adopt a similar perspective for estimating migration probabilities from data on "infant migration," which are obtained from age-specific population data cross-tabulated by current place of residence and place of birth. Children who are, say, 0–4 years old at the time of the census and living in region *j*, having been born in region *i*, must have migrated during the immediately preceding 5-year interval. Given their young age, and the fact that they were on average born two and a half years ago, it is unlikely that they experienced more than one migration. These data provide our initial estimate of age and spatial interaction and of migration level. Regression equations, model migration schedules, and log-linear models may be used to expand these population distributions into age-specific migration patterns.

To illustrate the regression method, consider the data presented in Fig. 6.5, which shows a plot of the aggregate conditional survivorship proportion, $S_{ij}(+)$, against the corresponding first age-group-specific component of that aggregate proportion, $S_{ij}(\text{-}5)$. The former represents the fraction of persons of all ages who resided in region *i* at the start of the time interval and in region *j* at the end of it. The latter is the first member of the set of age-group-specific proportions $S_{ij}(x)$, that in a suitably weighted linear combination comprise the former; it represents the fraction of all births born in region *i* during the past, say 5 years, who survived to the census date to enter the 0–4 years age group resident in region *j* at that date. Consequently, it

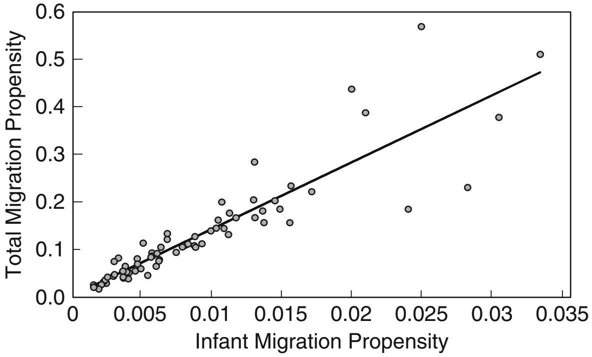


Fig. 6.5 Total migration propensity as a function of infant migration propensity: U.S. interdivisional migration. (Census 2000)

can be calculated by back-casting to region *i* all *i*-born 0–4 year olds enumerated at the time of the census, no matter where they lived, and then deriving the fraction of that number who ended up in region *j* at the time of the census count. (The S_{ij}(-5) measure is defined on pp. 98–99 of Rogers, 1995).

Examining the scatter plot in Fig. 6.5, we notice that a straight line offers a good approximation of the relationship between the infant migration level (S_{ij}(-5)) between regions *i* and *j* and the corresponding level across all ages ($R^2 = 0.84$). Separate regression equations need to be specified in order to estimate migration schedules with a retirement peak. And observed regularities in patterns of age-specific migration probabilities suggest that information on the probabilities of infant migration also can be linked to the corresponding probabilities in each of the subsequent age groups by means of a regression equation (Rogers & Jordan, 2004). We, therefore, can consider a linear regression that links each age-specific $S_{ij}(x)$ with S_{ij}(-5):

$$S_{ij}(x) = \beta_0 + \beta_1 S_{ij}(\text{-}5) + \varepsilon_{ij}, \tag{6.2}$$

where the βs are the parameters of the regression model and $\varepsilon_{ij}$is the error term. Using this simple linear regression equation, estimated migration propensities for each of the subsequent 5-year age cohorts can be determined.

The ability of this model to predict migration in subsequent time periods depends largely on the consistency of the $S_{ij}(x)$ to S_{ij}(-5) relationship over time. This relationship may be tested by plotting observed regression parameters. Figure 6.6 shows slope coefficient values resulting from Eq. (6.2) applied to the five census periods between 1960 and 2000. The slope coefficients vary similarly across all periods; the *intercept* values vary only slightly (not shown), between -0.0015 and 0.0020, and may be roughly approximated by a point lying within that range or zero.

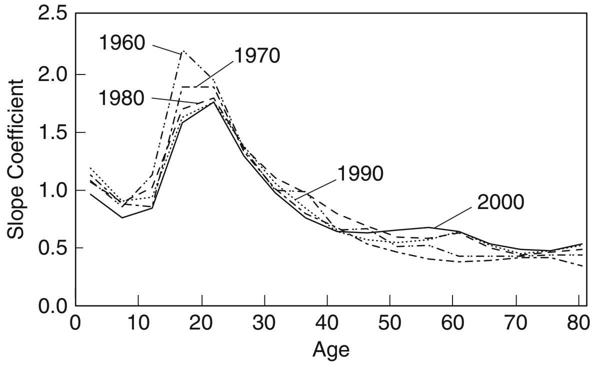


Fig. 6.6 Slope coefficients from the nine division simple linear regression model, 1960–2000

6.3.2 Estimates Based on Infant Migration Data: The Log-Linear Method

The loglinear-with-offset model can be thought of as a relational model (Rogers, Willekens et al., 2003). In this situation, the offset is the collection of 0–4 year old birthplace-specific population stocks. A log-linear-with-offset model can be specified which uses the 0–4 year old birthplace-specific population stocks to predict the aggregate patterns (assuming the marginal totals are known) of those aged 0–4 years at the time of the census, and effectively serves as a "proxy" for the interaction patterns of the current migration flows.

The log-linear-with-offset model can be used to include age-specific patterns. In this case, the offset contains structural zeros in the diagonal and the "migration" patterns of those aged 0–4 years at the time of the census in the off-diagonals. The overall age profile and aggregate proportions migrating from and to each region are assumed to be known. If instead one has to work with population totals, then one needs to estimate or borrow the aggregate age-specific proportions of migrants and non-migrants.

By way of illustration, consider the 0–4 year old "migration" patterns for U.S.-born persons set out in Table 6.4. The spatial structure of these "infant" migrants resembles that of the period migrants. The predicted aggregate flows from New England and South Atlantic are presented in Fig. 6.7. These predicted flows come from the log-linear model with two alternative offsets being used: (1) migrants only (where migrants represent the marginal totals) and (2) migrants and non-migrants (where population stocks represent the marginal totals). Although both models appear to predict the observed data well, the migrants-only model (not surprisingly) did considerably better. The likelihood ratio statistics for the two models were 132,799 and $-1{,}632{,}755$, respectively. The corresponding R^2 values were 0.99 and 0.96, respectively.

Table 6.4 The spatial structure of 0- to 4-year-old birthplace-specific population stocks in the U.S., 2000

	Destination									
Origin	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific	Total
(a) *Observed Flows (in thousands)*										
New England	0	14	5	2	17	2	3	3	6	52
Middle Atlantic	17	0	19	5	62	6	10	8	14	140
East North Central	5	15	0	25	39	24	17	15	18	158
West North Central	2	5	36	0	15	6	19	14	11	106
South Atlantic	12	42	44	13	0	35	30	15	28	219
East South Central	2	5	27	6	34	0	18	4	7	103
West South Central	4	8	20	19	33	15	0	23	26	147
Mountain	4	8	13	14	17	5	20	0	43	123
Pacific	8	15	25	18	42	10	32	65	0	214
Total	54	111	189	102	258	102	148	147	152	1,264
(b) *Multiplicative components*										
New England	0.000	3.126	0.678	0.500	1.572	0.464	0.505	0.469	0.915	0.041
Middle Atlantic	2.824	0.000	0.886	0.472	2.152	0.536	0.594	0.467	0.843	0.111
East North Central	0.708	1.063	0.000	1.918	1.204	1.903	0.929	0.835	0.958	0.125
West North Central	0.471	0.501	2.242	0.000	0.681	0.674	1.502	1.096	0.867	0.084
South Atlantic	1.317	2.166	1.344	0.743	0.000	1.974	1.160	0.594	1.063	0.173
East South Central	0.506	0.506	1.756	0.735	1.626	0.000	1.450	0.373	0.550	0.082
West South Central	0.587	0.653	0.920	1.591	1.099	1.231	0.000	1.328	1.438	0.117
Mountain	0.729	0.733	0.726	1.438	0.657	0.475	1.380	0.000	2.862	0.098
Pacific	0.882	0.779	0.771	1.017	0.967	0.559	1.277	2.617	0.000	0.170
Total	0.043	0.088	0.149	0.081	0.204	0.080	0.117	0.116	0.121	1,264

Source: Raymer and Rogers (2007, Table 6.5)

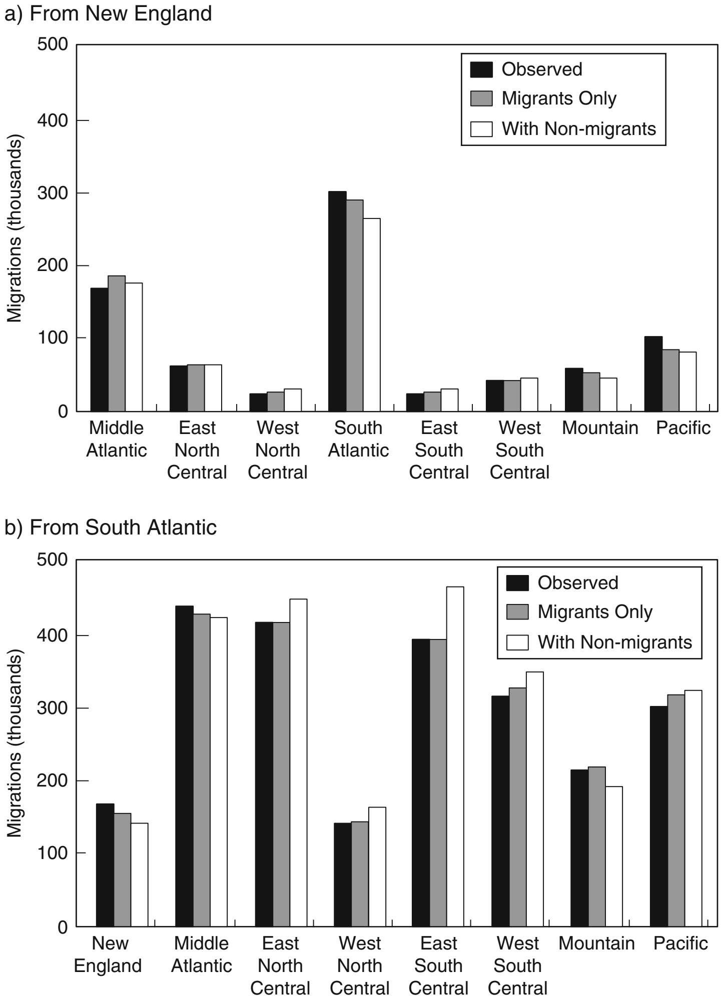


Fig. 6.7 A comparison of infant migration log-linear model predictions: U.S. interdivisional migration flows (in thousands) from New England and South Atlantic, 1995–2000 (*Source*: Raymer & Rogers, 2007, Fig. 6.8)

Age-specific predictions using log-linear models also did well, capturing the levels and most of the age profiles. Examples of such predictions are set out in Fig. 6.8. Our illustration applied a single age profile to estimate all age-specific patterns. The age profile is the same for both the migrants-only and the migrants and non-migrants models. This meant that the shapes of some flows, such as the retirement migration peak found in the Middle Atlantic to South Atlantic flow, were not entirely captured. For the flows set out in Fig. 6.8, the R^2 values were 0.88, 0.94, 0.97, and 0.95 for the New England-Middle Atlantic, Middle Atlantic-South Atlantic, South Atlantic-Middle Atlantic, and Pacific-South Atlantic flows. The corresponding likelihood

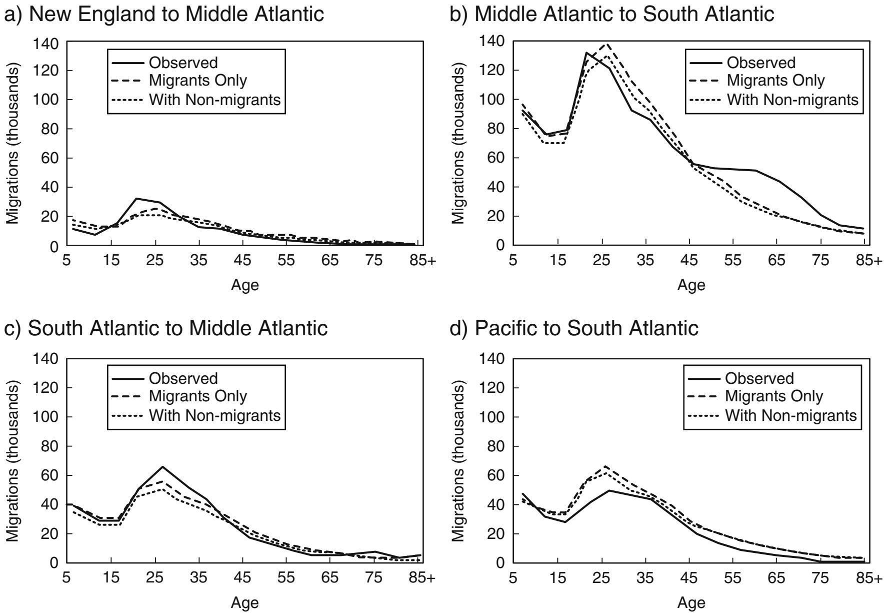


Fig. 6.8 A comparison of infant migration log-linear model predictions: Selected age-specific U.S. interdivisional migration flows (in thousands), 1995–2000. (*Source*: Raymer & Rogers, 2007, Fig. 6.9)

ratio statistics were lower for the migrants-only model, except for the Pacific-South Atlantic flow. Overall, the migrants-only model performed better.

6.3.3 Application to Historical Data

The above analysis focuses on data where the answers are known, which was necessary to test the model and to develop the modeling framework. In this section, we extend the analysis to an unknown situation. The aim is to estimate the 1905–1910 and 1915–1920 migration flows between the Northeast, Midwest, South and West regions based on birthplace-specific population stock information (of U.S.-borns) obtained from the 1900, 1910 and 1920 censuses.

The methodology we propose to estimate the 1905–1910 and 1915–1920 migration flows and conditional survivorship proportions is a simple one. We begin by interpolating linearly between the counts reported in the 1900, 1910, and 1920 censuses to obtain crude estimates of the 1905 and 1915 regional population totals. For example, our initial estimate of the 1905 population of the Northeast is (16,222,000 + 19,022,000)/2 = 17,622,000. For the Midwest it is 23,334,000, for the South it is 26,586,000, and for the West it is 4,308,000. Together, the four regional totals give us an estimated national population in 1905 of 71,851,000, with 25% in the Northeast, 32% in the Midwest, 37% in the South, and 6% in the

West. The next step is to use these relative shares to rescale the 1910 population of 78,279,000 to derive the number of survivors living in each of the four regions in 1905, resulting in 19,022,000 for the Northeast, 25,422,000 for the Midwest, 28,964,000 for the South, and 4,694,000 for the West.

Having obtained the column and row regional totals in this crude manner, we now have the marginal totals set out for 1905–1910 in Panel (a) of Table 6.5. The next step is to fill-in the missing elements of that matrix by imposing the spatial structure of the population aged 0–4 in 1910 that is presented in Panel (b) of Table 6.5. For example, 23,000 0-4-year olds residing in the Midwest in 1910 were reported to have been born in the Northeast. Hence, they are the surviving infant migrants during the 1905–1910 interval. To impose their spatial structure, we simply need to rescale their individual values to add up to the appropriate row and column totals described in Table 6.5 Panel (a). For New England that means raising the row total of 2,619,000 to 19,199,000, and so on for the other three regional totals. This first step, however, gives rise to column totals that do not match the ones set out in Panel (a) of Table 6.5. Consequently, the next step is to rescale the elements of each column to match the desired totals. But now the row elements no longer add up to the desired row totals. We therefore repeat our row and column rescalings until both the regional row totals and column totals match those presented in Table 6.5 Panel (a). This iterative process, known as biproportional adjustment (or iterative proportional fitting) was described in Chapter 3, Section 3.3. It took 703 iterations for the 1905–1910 estimated migration table to converge to predefined marginal row and column totals (note, the large number of iterations was necessary because the population totals were in the millions). In Table 6.5 Panel (c) exhibits the resulting "predicted" migration flows, and Panel (d) shows the associated conditional survivorship proportions.

Table 6.5 also presents parallel calculations for the 1915–1920 period. Examining these results, we find a number of interesting patterns. First, the highest out-migration flow in 1905–1910 was from the Midwest to the West, whereas the highest out-migration flow in 1915–1920 was the reverse flow from the West to the Midwest. All outflows from the Northeast were lower in 1915–1920 than were those a decade earlier. The reverse was the case for outflows from the West.

The crude estimation procedure described in this section is built on a number of assumptions that are easily challenged. First, the linear interpolation assumption that initiated the process could be replaced by a nonlinear one. Second, the foreign-born population and the impacts of international migration have been excluded. Our model is unable to estimate the internal migration patterns of groups born outside the country. And third, the spatial structure of 0-4-year old migrants is only an approximation of the corresponding structure of all age groups considered as an aggregate.

But a start has been made and, in the absence of migration data, one will always be faced with the uncertainties commonly associated with indirect estimation in general. Nevertheless, indirect estimation methods, such as the simple one proposed here, may help researchers to better understand the mechanisms of population change in the context of very limited information on migration.

Table 6.5 Indirect estimation of the 1905–1910 and 1915–1920 historical migration flows between regions in the U.S. (in thousands)

	1905–1910					1915–1920				
	Destination					Destination				
Origin	Northeast	Midwest	South	West	Total	Northeast	Midwest	South	West	Total
(a) Margins										
Northeast					19,199					22,591
Midwest					25,422					29,172
South					28,964					32,971
West					4,694					6,915
Total	19,022	24,929	28,770	5,557	78,279	22,864	29,159	32,361	7,264	91,648
(b) 0–4 year old birthplace-specific population stocks										
Northeast	2,576	23	14	6	2,619	2,997	30	18	4	3,049
Midwest	10	3,049	39	42	3,140	16	3,382	44	39	3,481
South	8	30	4,015	13	4,065	20	56	4,004	20	4,100
West	3	10	7	619	638	2	18	10	820	850
Total	2,597	3,112	4,074	680	10,462	3,035	3,485	4,076	883	11,479
(c) Predicted migration flows										
Northeast	18,860	171	84	83	19,199	22,346	151	72	21	22,591
Midwest	81	24,452	248	641	25,422	203	28,325	303	340	29,172
South	69	267	28,418	210	28,964	292	546	31,923	210	32,971
West	12	39	21	4,622	4,694	23	137	63	6,692	6,915
Total	19,022	24,929	28,770	5,557	78,279	22,864	29,159	32,361	7,264	91,648
(d) Predicted conditional survivorship proportions										
Northeast	0.9824	0.0089	0.0044	0.0043	1.0000	0.9892	0.0067	0.0032	0.0009	1.0000
Midwest	0.0032	0.9618	0.0097	0.0252	1.0000	0.0069	0.9710	0.0104	0.0117	1.0000
South	0.0024	0.0092	0.9811	0.0073	1.0000	0.0089	0.0165	0.9682	0.0064	1.0000
West	0.0026	0.0083	0.0044	0.9848	1.0000	0.0034	0.0198	0.0091	0.9677	1.0000

6.4 Inferring Current Spatial Patterns Using Combined Data Sets

As was illustrated in earlier chapters, the ACS PUMS sample data are not very effective at capturing detailed migration patterns, particularly for regions with small population totals. Thus far in this book, we have focused on improving the age patterns of migration data provided to the public by the ACS PUMS samples. This section focuses on the possibility of improving the spatial patterns of the ACS migration data by combining it with migration data provided by the IRS. The result is an enhanced data base that can be used for analysis or population planning.

An example of such an approach is described in a recent paper by Raymer et al. (2007), in which health registration data and census data are combined to predict detailed elderly migration flows in England and Wales over time. The health registration data provided the origin, destination, age and sex structures of the predicted migration flow tables. The census data provided the detailed structures, for example, migration by health status . A log-linear model predicts the flows and the census data represents the auxiliary information to be combined with the (incomplete) registration data. For our study here, the IRS migration data are used as an offset (or auxiliary information) in the following log-linear model, weighted to the marginal totals of the ACS migration data:

$$\ln \hat{n}_{ij} = \lambda + \lambda_i^O + \lambda_j^D + \ln\left(n_{ij}^*\right) \tag{6.3}$$

where n_{ij}^* denotes the IRS migration data. Total migrations between the thirteen states in the U.S. West Region during the 2004–2005 and 2005–2006 periods are used to illustrate the methodology. For comparison, the Census 2000 data are also examined.

6.4.1 Description of Migration Flows Collected from Different Data Sources

In seeking another data set to combine with the ACS, we looked for one that also asked a 1-year ago migration question and that could be constructed on an annual basis for the same years as the ACS data. The matching of U.S. Federal income tax returns by the IRS allows for the determination of both the origin and destination of migrants (up to the county level). It is thought that over 95% of the U.S. population is covered by this data series (Gross, 2005). The IRS makes these data available to the U.S. Census Bureau, after stripping away taxpayer names and Social Security numbers from the 1040 Individual Master File dataset. The Census Bureau geocodes these data, assigning a set of codes to each location-specific tax return, and uses these data to update its demographic database in between the decadal censuses.

Analysis of the IRS data reveals that the spatial patterns of *interstate* migration generally remain surprisingly stable over time, despite changing economic conditions (see, for example, Engels and Healy, 1981). This stability of the IRS data suggests its use to "repair" the irregularities in the ACS PUMS sample migration data.

In order to combine migration data from different sources in a study, one may have to first account for differences in measurement (Bell et al., 2002; Long & Boertlein, 1990; Morrison, Bryan, & Swanson, 2004; Rogers, Raymer et al., 2003; Rogerson, 1990; United Nations, 1992). For example, migration events, which can occur multiple times within a one year time period, are captured by population registration systems while changes in residential status (or transitions) from one point in time to another are captured by censuses (and surveys). These two data collection systems capture two different types of migration data, i.e., "migrations" and "migrants" (Rees and Willekens, 1986), but this does not necessarily prevent combining the data. For example, in comparing health registration and census data for England and Wales, Boden, Stillwell, and Rees (1992) found high levels of correlation between the in-migration, out-migration and net migration totals. More recently, Raymer et al. (2007), in analyzing elderly internal migration (also England and Wales), noted that the main differences between the 2000–2001 health registration flows and the 2001 Census flows were in the levels of migration. The spatial patterns, on the other hand, were very similar after controlling for the levels. Knowing that the census and population health registers have similar underlying structures allowed Boden et al. (1992) to combine these two sources to study the evolution of detailed migration patterns over time.

For the U.S., Engels and Healy (1981, p. 1354) found that annual IRS migration data, pooled from 1970–1973, were in "strong agreement" with the 5-year migration data obtained from the 1970 Census. We did not compare annual and 5-year migration as Engels and Healy did, but we borrowed their technique of using the Index of Dissimilarity (*D*) to examine the stability of annual IRS migration data for the years 2000, 2005 and 2006. In this context, this index measures changes in the distributions of migrant destinations (or origins) over two annual time periods, and for a specific state *j* it is defined as

$$D_j = 100\left(\frac{1}{2}\sum_{i=1}^{k}|x_{ij} - y_{ij}|\right) \tag{6.4}$$

where x_{ij} is the proportion of in-migrants (out-migrants) coming to (leaving from) state *j* from (to) state *i* during one year *n*, and y_{ij} is the proportion of in-migrants (out-migrants) coming to (leaving from) state *j* from (to) state *i* during another year *m*, and *k* is the number of destination (origin) states. The *D* values set out in Table 6.6 measure the "agreement" between the in- and out-migration patterns of the 2000, 2005, and 2006 IRS data. The closer *D* is to 100%, the more dissimilar the distributions are from perfect equality. The *D* values in Table 6.6 never exceed 11% and most values are significantly lower. Clearly, the spatial distributions of annual migration patterns were highly consistent during the 2000–2006 period.

Table 6.7 presents the Coefficient of Variation (*CV*) values for the same three years of IRS migration data: 2000, 2005, and 2006. Rogers and Raymer (1998)

Table 6.6 Indexes of dissimilarity (*D*) for IRS migration data comparing 2000, 2005 and 2006 in-migration and out-migration

	In-migration		Out-migration			In-migration		Out-migration	
State	2000–2005	2005–2006	2000–2005	2005–2006	State	2000–2005	2005–2006	2000–2005	2005–2006
Alabama	4.3	7.1	3.9	3.5	Missouri	4.3	3.3	4.3	2.7
Alaska	5.3	4.7	6.4	3.8	Montana	6.7	4.4	5.1	3.6
Arizona	10.4	2.6	5.7	3.7	Nebraska	5.0	4.9	4.5	3.3
Arkansas	5.6	7.0	3.9	4.1	Nevada	5.7	3.0	7.4	3.3
California	3.5	1.9	7.4	4.1	New Hampshire	5.9	4.6	9.1	5.2
Colorado	4.4	3.4	4.2	2.5	New Jersey	3.2	2.7	8.6	4.4
Connecticut	8.0	2.2	8.7	4.0	New Mexico	7.3	4.1	3.7	2.7
Delaware	4.7	4.2	4.9	3.1	New York	2.1	2.0	7.7	3.4
Wash D.C.	4.4	3.0	2.6	2.9	North Carolina	6.6	3.7	3.4	2.8
Florida	7.1	2.3	3.1	4.5	North Dakota	5.4	4.3	5.6	5.0
Georgia	5.0	7.5	3.3	2.4	Ohio	4.0	2.5	5.3	2.9
Hawaii	5.6	6.0	5.7	6.8	Oklahoma	3.8	4.2	3.2	2.5
Idaho	10.8	4.8	4.9	3.4	Oregon	8.8	2.4	5.0	2.8
Illinois	3.4	3.3	4.9	2.6	Pennsylvania	7.9	2.0	5.4	3.0
Indiana	5.4	3.3	4.6	3.1	Rhode Island	5.8	4.9	9.7	5.1
Iowa	4.3	3.0	5.1	3.2	South Carolina	4.7	3.9	3.2	2.4
Kansas	4.1	3.5	4.7	3.2	South Dakota	5.2	3.6	6.6	3.8
Kentucky	4.2	3.6	3.5	3.6	Tennessee	5.2	6.1	3.7	2.5
Louisianna	4.5	4.7	3.9	18.7	Texas	5.3	13.4	3.7	2.3
Maine	4.7	4.0	6.7	4.4	Utah	6.2	3.1	5.7	3.2
Maryland	4.0	2.5	6.9	3.8	Vermont	4.4	4.9	8.4	3.8
Massachusetts	3.9	2.5	9.1	3.5	Virginia	3.4	2.7	4.6	3.5
Michigan	4.0	3.2	5.0	3.0	Washington	5.1	4.3	4.8	3.0
Minnesota	4.4	3.4	5.4	2.6	West Virginia	7.8	4.4	7.2	2.8
Mississippi	5.1	12.4	4.2	6.0	Wisconsin	4.4	3.0	5.2	2.6
					Wyoming	5.4	4.2	6.1	3.5

Table 6.7 Coefficient of variation (*CV*) measures of IRS migration data: Distributions of in-migration and out-migration to and from each state, 2000, 2005 and 2006

	In-migration			Out-migration				In-migration			Out-migration		
State	2000	2005	2006	2000	2005	2006	State	2000	2005	2006	2000	2005	2006
Alabama	1.62	1.65	1.64	1.71	1.77	1.73	Missouri	1.42	1.43	1.38	1.37	1.38	1.39
Alaska	1.15	1.09	1.06	1.17	1.13	1.15	Montana	1.29	1.34	1.33	1.28	1.23	1.23
Arizona	1.57	2.15	2.19	1.51	1.36	1.34	Nebraska	1.40	1.40	1.36	1.34	1.31	1.34
Arkansas	1.62	1.57	1.56	1.66	1.66	1.70	Nevada	2.68	2.93	2.82	2.12	1.88	1.84
California	1.00	0.99	0.97	1.13	1.30	1.36	New Hamp.	2.82	3.16	2.94	2.10	1.94	1.96
Colorado	1.23	1.26	1.29	1.13	1.08	1.11	New Jersey	2.83	2.99	2.92	1.93	2.08	1.97
Connecticut	1.93	2.26	2.22	1.60	1.71	1.65	New Mexico	2.02	1.90	1.90	2.02	2.01	2.02
Delaware	2.35	2.40	2.40	2.10	2.10	2.05	New York	1.51	1.50	1.50	1.76	1.97	1.86
Wash. D.C.	2.94	2.78	2.72	3.71	3.73	3.64	N. Carolina	1.31	1.32	1.35	1.36	1.41	1.39
Florida	1.15	1.31	1.26	1.12	1.10	1.18	N. Dakota	2.02	2.17	2.19	2.22	2.24	2.18
Georgia	1.39	1.39	1.48	1.48	1.53	1.52	Ohio	1.12	1.13	1.13	1.22	1.28	1.25
Hawaii	1.77	1.90	1.70	1.71	1.62	1.53	Oklahoma	1.92	1.87	1.81	1.94	1.91	1.97
Idaho	1.87	2.09	2.21	1.76	1.75	1.77	Oregon	2.34	2.64	2.64	2.32	2.30	2.26
Illinois	1.12	1.14	1.13	1.17	1.24	1.23	Pennsylvania	1.57	1.83	1.84	1.39	1.48	1.44
Indiana	1.59	1.75	1.73	1.42	1.44	1.44	Rhode Island	2.24	2.53	2.44	2.11	2.16	2.11
Iowa	1.45	1.49	1.44	1.33	1.33	1.33	S. Carolina	1.63	1.60	1.58	1.85	1.83	1.83
Kansas	1.73	1.82	1.75	1.75	1.78	1.81	S. Dakota	1.42	1.43	1.38	1.38	1.29	1.30
Kentucky	1.53	1.50	1.53	1.57	1.59	1.55	Tennessee	1.19	1.18	1.24	1.28	1.31	1.28
Louisiana	1.85	1.78	1.92	2.09	2.05	3.07	Texas	1.05	1.18	1.64	0.98	0.99	0.97
Maine	1.57	1.71	1.65	1.48	1.53	1.48	Utah	1.67	1.84	1.89	1.47	1.42	1.39
Maryland	1.57	1.64	1.63	1.58	1.61	1.57	Vermont	1.63	1.70	1.69	1.61	1.60	1.56
Massachusetts	1.38	1.44	1.41	1.54	1.65	1.56	Virginia	1.22	1.20	1.19	1.27	1.34	1.34
Michigan	1.23	1.24	1.22	1.23	1.29	1.27	Washington	1.74	1.91	1.78	1.61	1.51	1.50
Minnesota	1.29	1.36	1.34	1.33	1.35	1.32	W. Virginia	1.82	1.95	1.98	1.84	1.78	1.77
Mississippi	1.73	1.69	2.11	1.58	1.63	1.55	Wisconsin	1.82	1.93	1.86	1.53	1.47	1.48
							Wyoming	1.34	1.31	1.27	1.38	1.37	1.38

showed that this index can be used to measure the degree of geographical concentration of the migration streams leaving (arriving in) a particular origin, in this case, states. When the *CV* is associated with the spatial structure of a state's in-migration (out-migration) stream, it is simply the standard deviation to mean ratio of the numbers of in-migrants (out-migrants) from (to) the other states. Comparing the geographical concentration of each state's in-migration and out-migration spatial patterns in each of the three years, we once again find considerable stability over time. For example, California's in-migration patterns exhibit the lowest *CV* values in each of the three data sets, and Texas's out-migration patterns do the same. The District of Columbia's out-migration pattern shows the highest *CV* values in all three data sets, and Oregon and North Dakota appear in the top three or four group all three times depending on whether we include Louisiana in 2006, in light of the Katrina disaster. For in-migration, the top group consists of the District of Columbia, New Jersey, Nevada, and New Hampshire in each of the three data sets; with the latter exhibiting the highest value in both IRS data sets. So, clearly, the spatial structures of the interstate migration streams in the United States tend to exhibit considerable stability over time, and migration data collected from different sources, in this case the decennial census and the IRS, tend to reflect this. We, therefore, explore this stability in further detail below.

The structure of the remaining section below is as follows. First, the ACS PUMS and the IRS interstate migration data for the 2005 and 2006 periods are described and compared. Second, after finding similar proportional structures in the migration tables, the IRS data are used as an offset in a log-linear model to infer (or improve) the spatial patterns of the ACS 2005 and 2006 PUMS data. As with the case of the earlier application in Section 6.3, age structures are ignored, and would need to be introduced with model migration proportion schedules, such as those discussed in Section 2.4.3.

6.4.2 Comparison of Migration Flows Collected from Different Data Sources

The proportions of migration flows from and to the thirteen states in the U.S. West region are set out in Fig. 6.9 for data obtained from Census 2000, the ACS 2005 and 2006 PUMS, and the 2005 and 2006 IRS. The basic conclusion from this comparison is that, although the three data sources measure migration differently, they all offer very similar descriptions of the in-migration and out-migration proportions to each state. This suggests that we can combine the ACS PUMS and IRS data together using the model presented in Eq. (6.3).

In Fig. 6.10, we continue the comparison but only show the proportions of migration to each destination from California (i.e., a large-sized population state), Washington (i.e., a medium-sized population state) and Wyoming (i.e., a small-sized population state). Here we see that the ACS PUMS sample migration data vary considerably more than the Census and IRS data. The patterns for Wyoming are

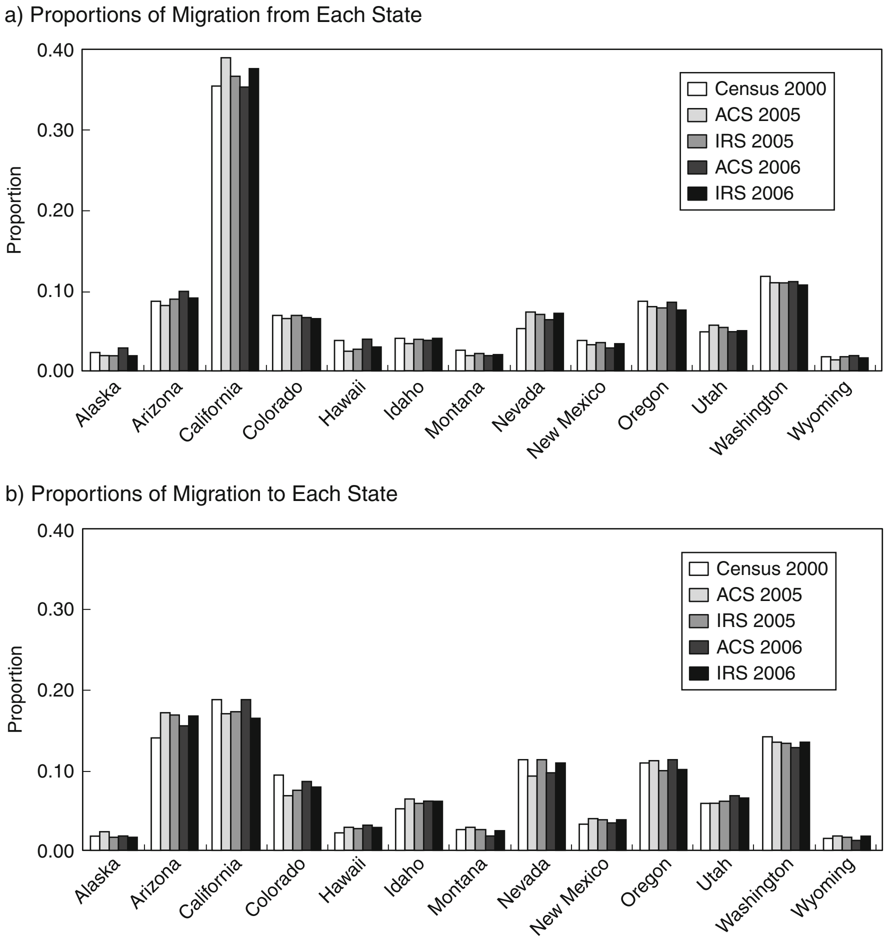
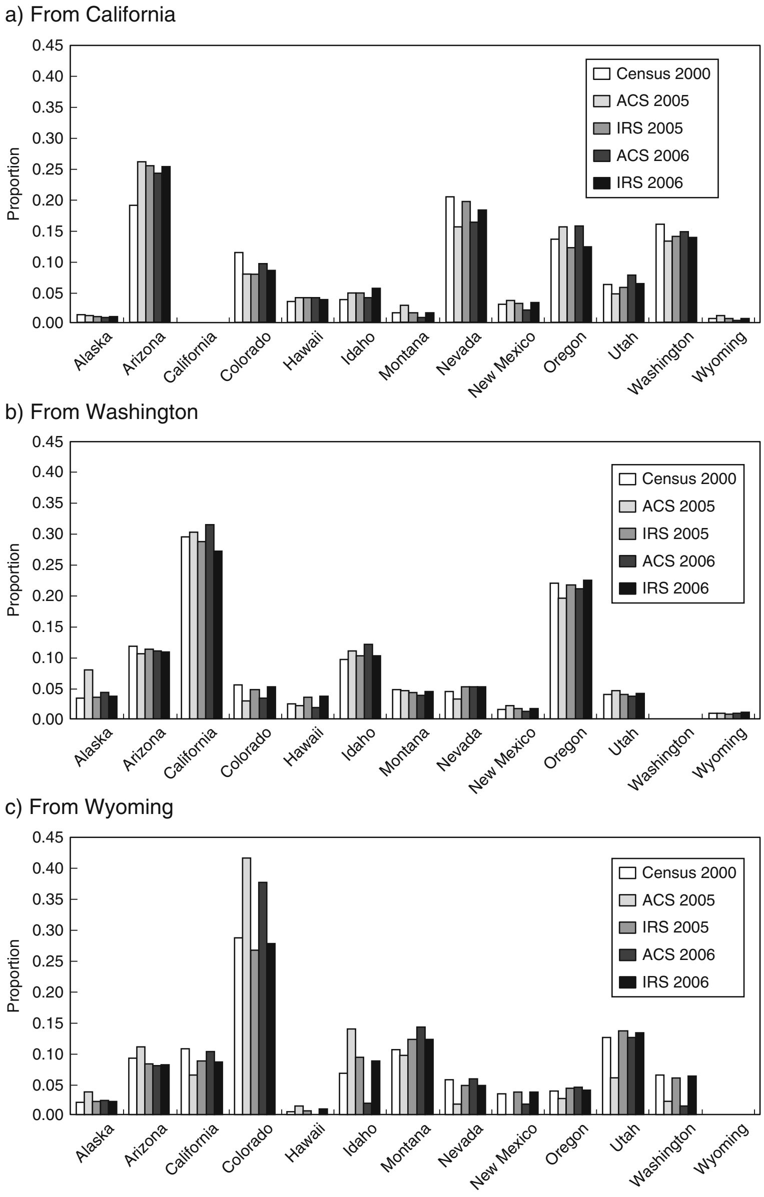


Fig. 6.9 A comparison of Census, ACS PUMS, and IRS migration flow data: Proportions of migration from and to each state in the U.S. West Region

particularly variable from 2005 to 2006. For example, the proportions of migration from Wyoming to Colorado are much higher in the ACS than they are in the Census or IRS data. Relative to the Census and IRS data, the proportion of migration from Wyoming to Idaho is much higher in the ACS 2005 but much lower in the ACS 2006. The ACS proportions to Utah in 2005 and Washington in 2005 and 2006 appear to be too low. Finally, the ACS PUMS data did not capture any migration at all between Wyoming and Hawaii in 2006 and between Wyoming and New Mexico in 2005, whereas the Census and IRS data did. Our conclusion from this analysis is that the IRS data appear to capture the spatial patterns of migration better than the ACS PUMS sample data, justifying its inclusion as an offset in the model presented in Eq. (6.3).



a) From California

b) From Washington

c) From Wyoming

Fig. 6.10 A comparison of Census, ACS PUMS and IRS migration flow data: Proportions of migration from California, Washington, and Wyoming to other states in the U.S. West Region

6.4.3 Prediction of ACS 2005 and 2006 Spatial Patterns Using IRS Data

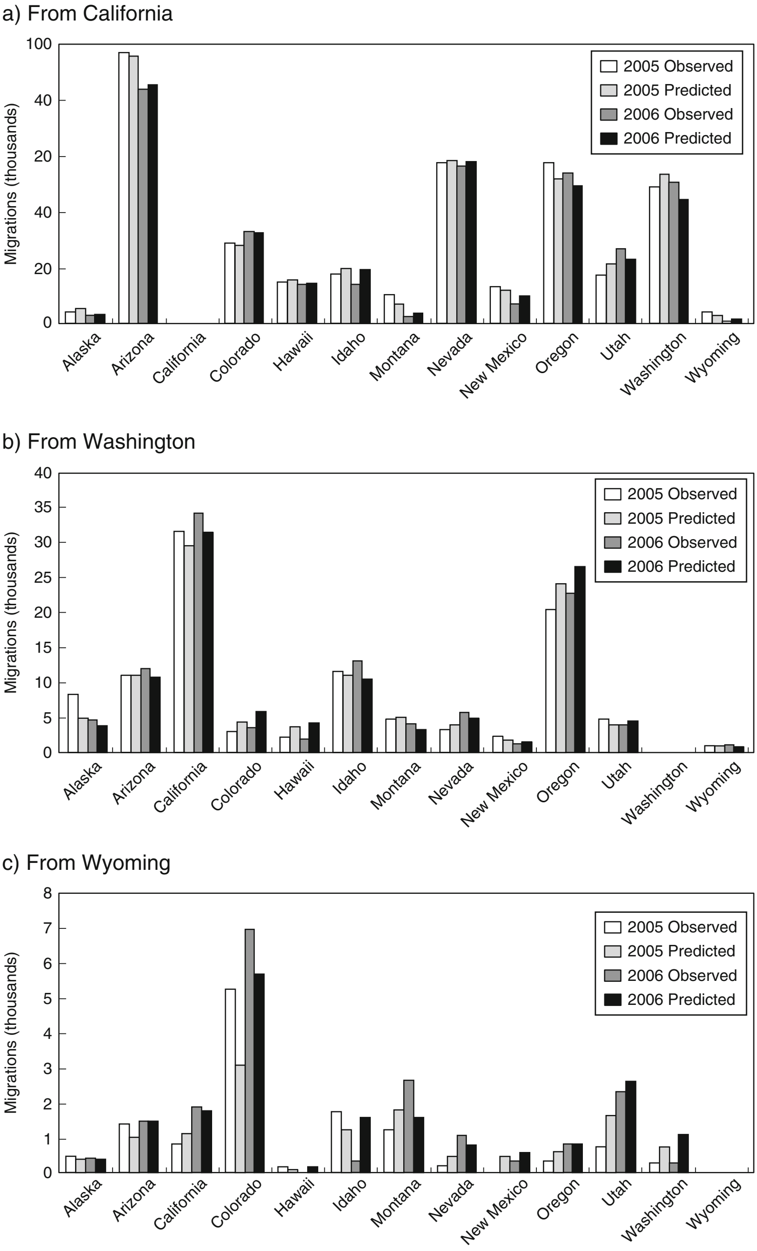
The results of applying the model in Eq. (6.3) to predict migration flows between states in the U.S. West region for the years 2005 and 2006 are set out in Fig. 6.11 for flows from California, Washington and Wyoming. The combination of IRS data with ACS PUMS data produced flows that were smoother and more realistic than those described by the original ACS data (e.g., see the Washington to Alaska or Wyoming to Nevada flows), particularly for the small state of Wyoming. Also, there are no longer any "missing" or zero flows, for example, in the Wyoming to Hawaii or Wyoming to New Mexico flows. In some cases, the predicted flows resulted in different patterns over time than those reported from the ACS PUMS data. For example, the ACS data exhibited decreases over time in the Wyoming to Idaho and Wyoming to Montana flows, whereas the predicted patterns resulted in increases over time.

For a more detailed comparison of the ACS PUMS data and predicted migration flows, consider the ratios of predicted to observed flows set out in Table 6.8. These values allow us to compare how much the predicted migration flows differed from the reported ACS values. Notice that the marginal totals of the reported ACS values and of the predicted values match perfectly. There are many cases where the ACS data and predicted flows are fairly similar, such as the flows to California. However, there are also some notable differences in the origin-destination patterns, such as the 2005 predicted flow from Montana to Alaska, which was predicted to be nearly five times greater than the corresponding ACS value. The largest predicted increases are the 2005 flow from Alaska to Utah (10.07 times) and the 2006 flow from Nevada to Alaska (16.37 times). As far as predicted reductions, the biggest differences were found in the 2005 Idaho to Hawaii flow (0.21 of the ACS flow) and the 2006 Alaska to Hawaii flow (0.32 of the ACS flow).

6.5 Summary and Discussion

This chapter has focused on the indirect estimation problem posed by a total absence of any observed migration data. Three different approaches to the problem have been explored. The first seeks to predict the age composition of total migration outflows leaving a region from information about the age composition of the population of that region. The second and third go further and strive, in addition, to include the regions of destination of each estimated outflow.

Rogers (1978a) and Castro and Rogers (1983) first showed that the age distributions of migration rates are related to the age distributions of the origin populations. In Little and Rogers (2007) and in Section 6.2.2 of this chapter, we take this finding a step further and demonstrate that by linking this association with a typology of model migration schedules that collectively describe the age composition of a large number of conforming schedules, allows one, in most instances, to infer the "family"



a) From California

b) From Washington

c) From Wyoming

Fig. 6.11 Observed and predicted ACS 2005 and 2006 PUMS flows from California, Washington, and Wyoming to other states in the U.S. West Region

Table 6.8 Ratios of predicted to observed ACS 2005 and 2006 PUMS sample data: Migration flows between states in the U.S. West Region

	Destination													
Origin	AK	AZ	CA	CO	HI	ID	MT	NV	NM	OR	UT	WA	WY	Total
(a) 2005														
Alaska		1.90	0.66	0.62	3.79	0.64	1.34	1.93	1.00	1.50	10.07	0.86	5.49	1.00
Arizona	1.02		1.14	1.44	0.70	1.00	4.58	0.83	1.45	1.01	0.70	0.56	0.73	1.00
California	1.35	0.98		0.97	1.05	1.13	0.69	1.02	0.91	0.90	1.22	1.09	0.69	1.00
Colorado	1.18	0.98	1.22		0.64	0.92	1.02	1.18	0.70	0.78	1.02	0.98	1.36	1.00
Hawaii	1.52	1.26	0.91	1.78		1.94	**NA**	1.53	0.39	1.26	3.11	0.72	0.51	1.00
Idaho	1.83	0.96	1.16	0.75	0.21		3.46	1.39	2.00	0.79	1.04	0.94	0.82	1.00
Montana	4.48	0.79	1.19	0.90	2.88	1.06		0.50	1.62	0.75	1.16	0.88	1.65	1.00
New Mexico	1.40	1.03	0.92	1.10	1.16	1.52	1.48		1.19	1.64	0.74	1.04	0.88	1.00
Nevada	2.46	0.97	0.99	0.96	2.19	0.53	0.53	0.82		6.82	0.92	1.37	0.95	1.00
Oregon	1.24	1.09	0.98	1.38	1.01	0.93	0.57	0.67	1.02		0.85	1.12	0.54	1.00
Utah	0.52	1.13	0.99	0.87	0.57	0.87	2.22	0.89	1.94	1.39		1.00	1.35	1.00
Washington	0.59	1.00	0.93	1.46	1.67	0.96	1.05	1.20	0.76	1.18	0.82		0.99	1.00
Wyoming	0.84	0.73	1.38	0.58	0.46	0.71	1.47	2.24	**NA**	1.81	2.21	2.75		1.00
Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6.8 (continued)

	Destination												
Origin	AK	AZ	CA	CO	HI	ID	MT	NV	NM	OR	UT	WA	WY	Total
(b) 2006														
Alaska		0.64	0.74	1.47	0.32	2.60	5.36	0.69	0.77	1.87	1.67	1.57	1.28	1.00
Arizona	1.41		1.14	0.76	0.83	1.06	1.22	1.17	1.06	0.82	0.89	1.01	0.92	1.00
California	1.13	1.02		1.00	1.04	1.37	1.45	1.03	1.45	0.91	0.87	0.88	1.56	1.00
Colorado	1.04	1.36	1.16		0.71	0.51	1.18	0.57	0.88	0.95	1.49	1.10	0.70	1.00
Hawaii	0.47	0.71	1.12	1.27		0.47	0.29	1.21	0.47	0.71	1.41	1.36	**NA**	1.00
Idaho	2.06	0.48	0.93	2.23	**NA**		1.10	0.87	1.33	1.58	1.04	0.86	2.35	1.00
Montana	0.96	0.50	2.10	1.24	0.35	0.52		2.93	3.33	1.30	1.23	1.02	1.45	1.00
New Mexico	0.65	1.17	0.88	1.45	0.99	1.71	1.80		0.63	0.99	1.22	0.82	7.71	1.00
Nevada	16.37	1.05	1.01	0.88	0.41	3.41	1.09	0.87		0.85	1.54	1.27	0.57	1.00
Oregon	1.03	1.12	0.88	0.94	2.38	0.81	0.89	2.13	0.64		0.71	1.14	0.51	1.00
Utah	3.34	1.13	1.15	1.00	2.17	0.93	1.01	0.88	0.47	0.93		0.83	1.30	1.00
Washington	0.82	0.90	0.92	1.64	2.17	0.80	0.79	0.86	1.18	1.16	1.14		0.73	1.00
Wyoming	0.93	0.99	0.93	0.82	**NA**	4.69	0.60	0.75	1.84	1.00	1.13	4.05		1.00
Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

memberships of unobserved age compositions of migrants, and, if given an estimated total number of out-migrants leaving a region, predict the likely age-specific number of out-migrants from that region.

The second approach to resolving the missing data estimation problem is addressed by using observed place-of-birth data, in addition to the observed place-of-residence data. The spatial migration pattern of infants, or of young children under five, underlying such data provides us with an initial estimate of age, spatial interaction, and migration level. This initial estimate can then be modified and expanded by using a combination of regression equations, model migration schedules, and log-linear models.

Finally, the third approach to the missing data problem follows the "combined data" strategy recommended by Willekens (1994), in which data from several sources are merged, integrated, and forced with log-linear models to be internally consistent. In Section 6.4 of this chapter, we illustrate our approach using the particular combination of ACS data on population stocks with IRS data on the migration of income tax filers.

All three methods show promise and merit further study. First, the age composition of origin populations and their out-migrants clearly are linked and therefore suggest an estimation strategy that seeks to capitalize on this association. Second, age profile regularities in counts and propensities of out-migration suggest that, inasmuch as children migrate as part of families, an estimation strategy that links the two should be a promising start. Third, combining data from different sources, with at least one of them reporting a change of usual residence, suggest yet another estimation strategy that works well.

But, most likely, additional information would improve the indirect estimation results. For example, introducing covariates, i.e., variables that change in parallel with changes in migration patterns, should be a profitable strategy. More research on this aspect of the indirect estimation of migration is therefore warranted.

Finally, if combining data from several different sources is a promising avenue for developing indirect estimation methods, then it is likely that combining several different estimation strategies should also be an equally promising overall strategy. For example, an appropriate "marriage" of the three approaches described in this chapter surely would produce better results than those produced by any one of the three used alone.

Chapter 7
Conclusion

Mortality, fertility, and migration rates combine to shape the temporal evolution of multiregional populations, and demographers study how the levels, age profiles, and spatial patterns of these contributors to population change vary over time and space. What they have discovered is that all three generally exhibit persistent regularities in their age and spatial structures, when changing levels are controlled for. Drawing on such regularities, it is often possible to improve the quality of the available data by smoothing irregular data, imposing the structures of borrowed and related data on inadequate data, or by inferring missing data.

The United Nations has been at the forefront of assembling and testing alternative methods for improving demographic data. Its now somewhat outdated manuals on this subject quickly became classic reports on the state of the art (United Nations, 1967, 1983). But they totally ignored the problem of estimating migration from incomplete data. More recently, a chapter on indirect estimation methods in an important text on formal demography (Preston, Heuveline, & Guillot, 2001) also totally ignores the case of migration. Demographic texts that do include topics on migration estimation tend to focus on residual methods (eg., Rowland, 2003; Siegel & Swanson, 2004). The indirect estimation of fertility and mortality has a long history in demography. A common strategy there has been to combine empirical regularities with other information to fill in the missing data. The indirect estimation of migration flows has a briefer history, in part because the estimation task is more complicated. The age pattern of migrants depends on the directions of migration. To be effective, therefore, a method must somehow integrate the age pattern with the corresponding spatial pattern. Nonetheless, efforts to indirectly estimate migration streams continues (Hill, 1985; Nair, 1985; Schmertmann, 1992; Warren & Kraly, 1985; Warren & Peck, 1980; Willekens, 1999; Zaba, 1987), notably those attempting to infer international or undocumented flows. This book adds to that literature by providing a manual for estimating age and origin-destination-specific migration flows of migrants in situations where the available migration data are suspect or missing. Although our focus has been on internal migrants, recent research has shown the methods also to be applicable to international migration (Raymer & Willekens, 2008).

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1_7,

Migration flow patterns exhibit strong age and spatial regularities. In a discussion of new “laws” of migration, Tobler (1995, pp. 335–337) argued that one of the most studied regularities is the age profile of migrants. He then focused on spatial patters of migrants and presented evidence of the correlation between six U.S. state-to-state tables for the contiguous United States. A deeper analytical examination of this issue appears in this book’s formal definitions of the age and spatial structures of migration, and how they can be represented by model migration schedules and multiplicative log-linear modeling frameworks (Rogers & Castro, 1986; Rogers, Willekens & Raymer, 2001, 2002; Rogers et al., 2002).

Our approach to improve the quality and reliability of observed data that are irregular has been to first smooth it. When the data are still inadequate and exhibit obviously unrealistic rates, and smoothing these does not resolve the problem, then our approach imposes data structures borrowed from elsewhere to repair the suspect data. Finally, when no migration data are available, we infer the rates using data on birthplace-specific counts of population stocks or adopt changes of address data, collected by other government agencies for other purposes (e.g., tax return data).

Underlying our efforts to smooth irregular data, to impose borrowed data onto inadequate data, or to infer structures exhibited in flows when observed data are missing all rely basically on two kinds of models: models of age patterns (model migration schedules) and models of spatial patterns (log-linear models). The former are described in Chapter 2; the latter are reviewed in Chapter 3. These models are used in Chapters 4, 5, and 6 to deal, respectively, with smoothing, imposing, and inferring strategies. In this concluding chapter, we touch on a few issues and approaches not covered in the book and deserving further research.

Several directions of further study are evident. First of all, how stable are the observed age and spatial patterns over time? Our preliminary examination of temporal stability suggests that many flows continue to exhibit a surprising degree of constancy in age profiles and spatial patterns. Rogers and Raymer (1999) found evidence of temporal stability in the interregional migration patterns of the foreign-born in the United States reported by the 1960, 1970, 1980 and 1990 censuses. More recently, Raymer, Bonaguidi, and Valentini (2006) showed strong regularities in the age patterns of Italian interregional migration from 1970 to 2001, with a gradual “ageing” of the labor force peak. But more definitive findings are needed.

Second, our focus has been directed at internal migration age patterns. What about international migration patterns? Rogers and Raymer (1999), for example, found that US immigration age profiles conformed to the Rogers-Castro model schedule during the 1955–1990 time period. Similar evidence for population movements in the European Union is presented in Chapter 10 of Raymer and Willekens (2008). But, once again, more definitive findings are needed.

Third, what about migration age and spatial patterns in the less developed countries? Evidence gathered to date show that model migration schedules and log-linear models also describe migration patterns in the less developed world. Muhidin (2002: 272–276) and Rogers, Jones, Partida, and Muhidin (2007) present supporting evidence for Indonesia, and the latter do the same for Mexico. It is likely, therefore, that patterns of migration in less developed countries also exhibit regularities, particularly of spatial patterns. But further studies are warranted.

Fourth, additional status categories besides age and sex can be introduced into the log-linear formulation, for example, ethnicity and employment. Smith et al. (2010) offer an illustration.

Fifth, we have examined migration data that reported 5-year (census) and 1-year (ACS) time interval flows. The so-called "1 year/5 year" problem was not studied, yet a method for transforming one into the other, however crude, would be useful. An effort to do this appears in Rogers, Raymer, et al. (2003), but it has not been carried far enough to merit inclusion in this book.

Sixth, censuses or surveys that produce distributions of persons cross-classified by age, *birthplace*, and place of current residence for two consecutive points in time offer information that can be used to infer the migration patterns that helped to shape such distributions. We have explored two of these in a pair of articles (Rogers & Raymer, 2005; Rogers & Liu, 2005, respectively). The first approach adopts observed past regularities in the relative intensities of *secondary* (return and onward) versus *primary* migration streams to indirectly estimate migrant flows. Originally proposed and tested by Ledent (personal communication, 2003) this approach was less than moderately successful and so was not included in this book.

The second approach, originally proposed by Castro (1985), decomposes estimated net migrant flows into the underlying gross in- and out- migrant streams to indirectly estimate migration flows. Here too, the approach encountered problems that led us to omit it from this book. Perhaps what *is* needed is a mixing of these approaches.

Finally, as we noted earlier, the indirect estimation of migration flows in the absence of observed migration data is similar to a paleontologist's effort to reconstruct an entire dinosaur from a newly discovered hip bone. In both cases, assessments of accuracy are generally problematic. In its 1983 manual on indirect estimation, the United Nations recognizes this problem:

> Perhaps the most serious limitation of Manual X is that it does not provide sufficient guidance for the assessment of results, an aspect of analysis that is also somewhat perfunctorily considered in the literature (United Nations,1983, p. 1).

Assessments of alternative indirect estimates may be carried out in much the same way as are assessments of alternative population forecasts. In both endeavors the results cannot be compared to the corresponding "true" numbers, except in tests carried out with past observed data. One therefore has to assess the quality of the procedures used to produce the desired numbers and to examine the "reasonableness" of the results. The assessment suggests further improvements and the procedures may be reapplied until a satisfactory solution is found. Research on this topic is warranted.

In the last paragraph of the final chapter of his textbook on multiregional demography, Rogers (1995) concludes:

> Multiregional demography is a relatively young branch of formal demography, and its potential contributions are becoming more widely recognized. Further progress in the field

> will depend to a large extent on the availability of the necessary detailed data for carrying out the analyses and projections that would promote its further development and acceptance. (Rogers, 1995, p. 208).

This book was written with the goal of expanding the possibilities of increasing the availability of adequate migration data in mind. Much more remains to be done, but it is hoped that a useful start has been made.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Andersson, A. E., & Holmberg, I. (1980). *Migration and settlement: 3. Sweden*. Research report 80-5. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Aufhauser, E., & Fischer, M. M. (1985). Log-linear modelling and spatial analysis. *Environment and Planning A, 17*(7), 931–951.

Bacharach, M. (1970). *Biproportional matrices and input-output change*. Cambridge: Cambridge University Press.

Bates, J., & Bracken, I. (1982). Estimation of migration profiles in England and Wales. *Environment and Planning A, 14*(7), 889–900.

Bates, J., & Bracken, I. (1987). Migration age profiles for local-authority areas in England, 1971–1981. *Environment and Planning A, 19*(4), 521–535.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., et al. (2002). Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society. Series A, 165*(3), 435–464.

Bennett, R. J., & Haining, R. P. (1985). Spatial structure and spatial interaction: Modeling approaches to the statistical analysis of geographical data. *Journal of the Royal Statistical Society. Series A, 148*(1), 1–36.

Boden, P., Stillwell, J., & Rees, P. (1992). How good are the NHSCR data? In J. Stillwell, P. Rees & P. Boden (Eds.), *Migration processes and patterns: Population redistribution in the United Kingdom* (Vol. 2, pp. 13–27). London: Belhaven Press.

Bogue, D. J., Hinze, K. E., & White, M. J. (1982). *Techniques of estimating net migration*. Chicago, Illinois: Community and Family Study Center, University of Chicago.

Brass, W. (1974). Perspectives in population predictions: Illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society A, 134*(4), 532–570.

Brown, K. M., & Dennis, J. E., Jr. (1972). Derivative free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares approximation. *Numerische Mathematik, 18*(4), 289–297.

Castro, L. J. (1985). *Analysis of age-specific gross and net migration schedules*. Paper presented at the 1985 Annual Meeting of the Population Association of America Boston, Massachusetts.

Castro, L. J., & Rogers, A. (1981). *Model migration schedules: A simplified formulation and an alternative parameter estimation method.* (Working paper 81-63). Laxenburg, Austria: International Institute for Applied Systems Analysis.

Castro, L. J., & Rogers, A. (1983). What the age composition of migrants can tell us. *Population Bulletin of the United Nations, 15*, 63–79.

Citro, C. F. (1998). Model-based small-area estimates: The next major advance for the federal statistical system for the 21st century. *Chance, 11*(3), 40–41, 50.

Clayton, C. (1977). Structure of interstate and interregional migration: 1965–1970. *Annals of Regional Science, 11*(1), 109–122.

A. Rogers et al., *The Indirect Estimation of Migration*, The Springer Series on Demographic Methods and Population Analysis 26, DOI 10.1007/978-90-481-8915-1,

Coale, A., & Trussell, J. (1996). The development and use of demographic models. *Population Studies, 50*(3), 469–484.

Coale, A. J., & McNeil, D. R. (1972). Distribution by age of frequency of first marriage in a female cohort. *Journal of the American Statistical Association, 67*(340), 743–749.

Drewe, P., & Willekens, F. (1980). Maximum likelihood estimates of age-specific migration flows in the Netherlands. *Delft Progress Report, 5*, 92–111.

Engels, R. A., & Healy, M. K. (1981). Measuring interstate migration flow: An origin destination network based on Internal Revenue Service records. *Environment and Planning A, 13*(11), 1345–1360.

Franklin, R. S., & Plane, D. A. (2006). Pandora's box: The potential and peril of migration data from the American Community Survey. *International Regional Science Review, 29*(3), 231–246.

George, M. V. (1994). *Population projections for Canada, provinces and territories, 1993–2016*. Ottawa: Statistics Canada, Demography Division, Population Projections Section.

Good, I. J. (1963). Maximum-entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics, 34*(3), 911–934.

Griffin, D. H., & Waite, P. J. (2006). American Community Survey overview and the role of external evaluations. *Population Research and Policy Review, 25*(3), 201–223.

Gross, E. (2005). http://www.irs.gov/pub/irs_soi/05gross.pdf

Gumbel, E. J. (1941). The return period of flood flows. *Annals of Mathematical Statistics, 12*(2), 163–190.

Haynes, K. E., & Fotheringham, A. S. (1984). *Gravity and spatial interaction models.* Beverly Hills, CA: Sage.

Heligman, L., & Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries, 107*(434), 49–80.

Hill, K. (1985). Indirect approaches to assessing stocks and flows of migrants. In D. B. Levine, K. Hill, & R. Warren (Eds.), *Immigration statistics: A story of neglect* (pp. 205–224). Washington, DC: National Academy Press.

Hofmeyr, B. E. (1988). Application of a mathematical model to South African migration data, 1975–1980. *Southern African Journal of Demography, 2*(1), 24–28.

Isard, W. (1960). *Methods of regional analysis: An introduction to regional science*. Cambridge: MIT Press.

Kawabe, H. (1990). *Migration rates by age group and migration patterns: Application of Rogers' migration schedule model to Japan, The Republic of Korea, and Thailand*. Tokyo: Institute of Developing Economies.

Kimball, B. F. (1956). The bias in certain estimates of the parameters of the extreme-value distribution. *Annals of Mathematical Statistics, 27*(3), 758–767.

Kitsul, P., & Philipov, D. (1982). High- and low-intensity model of mobility. In K. C. Land & A. Rogers (Eds.), *Multidimensional mathematical demography* (pp. 505–534). New York: Academic Press.

Knudsen, D. C. (1992). Generalizing Poisson regression: Including a priori information using the method of offsets. *Professional Geographer, 44*(2), 202–208.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quarterly of Applied Mathematics, 2*, 164–168.

Liaw, K.-L., & Nagnur, D. N. (1985). Characterization of metropolitan and nonmetropolitan out-migration schedules of the Canadian population system, 1971–1976. *Canadian Studies in Population, 12*(1), 81–102.

Lin, G. (1999a). Assessing changes in interstate migration patterns of the United States elderly population, 1965–1990. *International Journal of Population Geography, 5*(6), 411–424.

Lin, G. (1999b). Assessing structural change in U.S. migration patterns: A log-rate modeling approach. *Mathematical Population Studies, 7*(3), 217–237.

Little, J. S., & Rogers, A. (2007). What can the age composition of a population tell us about the age composition of its out-migrants? *Population Space and Place, 13*(1), 23–39.

Long, J. F., & Boertlein, C. G. (1990). *Comparing migration measures having different intervals. Current Population Reports, Series P-23, No. 166, pp. 1–11*. Washington, DC: U.S. Census Bureau.

Manson, G. A., & Groop, R. E. (1996). Ebbs and flows in recent U.S. interstate migration. *Professional Geographer, 48*(2), 156–166.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics, 11*(2), 431–441.

Mather, M., Rivers, K. L., & Jacobsen, L. A. (2005). The American community survey. *Population Bulletin, 60*(3), 3–20.

McNeil, D. R., Trussell, T. J., & Turner, J. C. (1977). Spline interpolation of demographic data. *Demography, 14*(2), 245–252.

Morrison, P. A., Bryan, T. M., & Swanson, D. A. (2004). Internal migration and short-distance mobility. In J. S. Siegel & D. A. Swanson (Eds.), *The methods and materials of demography* (pp. 493–521). San Diego: Elsevier Academic Press.

Mueser, P. (1989). The spatial structure of migration: An analysis of flows between states in the USA over three decades. *Regional Studies, 23*(3), 185–200.

Muhidin, S. (2002). *The population of Indonesia: Regional demographic scenarios using a multiregional method and multiple data sources*. Amsterdam: Rozenberg Publishers.

Nair, P. S. (1985). Estimation of period-specific gross migration flows from limited data: Bi-proportional adjustment approach. *Demography, 22*(1), 133–142.

O'Brien, L. (1992). *Introducing quantitative geography: Measurement, methods and generalised linear models*. London: Routledge.

Plane, D. A. (1984). A systemic demographic efficiency analysis of United States interstate population exchange, 1935–1980. *Economic Geography, 60*(4), 294–312.

Plane, D. A., & Mulligan, G. F. (1997). Measuring spatial focusing in a migration system. *Demography, 34*(2), 251–262.

Potrykowska, A. (1986). Modelling inter-regional migrations in Poland, 1977–1981. *Papers of the Regional Science Association, 60*, 29–40.

Potrykowska, A. (1988). Age patterns and model migration schedules in Poland. *Geographia Polonica, 54*, 63–80.

Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modelling population processes.* Oxford, UK: Blackwell.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Raymer, J. (2007). The estimation of international migration flows: A general technique focused on the origin-destination association structure. *Environment and Planning A, 39*(4), 985–995.

Raymer, J., Abel, G., & Smith, P. W. F. (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society. Series A, 170*(4), 891–908.

Raymer, J., Bonaguidi, A., & Valentini, A. (2006). Describing and projecting the age and spatial structures of interregional migration in Italy. *Population Space and Place, 12*(5), 371–388.

Raymer, J., & Rogers, A. (2007). Using age and spatial flow structures in the indirect estimation of migration streams. *Demography, 44*(2), 199–223.

Raymer, J., & Rogers, A. (2008). Applying model migration schedules to represent age-specific migration flows. In J. Raymer & F. Willekens (Eds.), *International migration in Europe: Data, models and estimates* (pp. 175–192). Chichester: Wiley.

Raymer, J., & Willekens, F. (Eds.). (2008). *International migration in Europe: Data, models and estimates.* Chichester: Wiley.

Rees, P. H. (1977). The measurement of migration, from census data and other sources. *Environment and Planning A, 9*(3), 247–272.

Rees, P. H., & Willekens, F. J. (1986). Data and accounts. In A. Rogers & F. J. Willekens (Eds.), *Migration and settlement: A multiregional comparative study* (pp. 19–58). Dordrecht: D. Reidel.

Robinson, G. (2009). *How to deal with estimates with low reliability*. Paper presented at the Annual meeting of the Population Association of America, Detroit, Michigan.

Rogers, A. (1967). Estimating interregional population and migration operators from interregional population distributions. *Demography, 4*(2), 515–531.

Rogers, A. (1968). *Matrix analysis of interregional population growth and distribution*. Berkeley: University of California Press.

Rogers, A. (1973). Estimating internal migration from incomplete data using model multiregional life tables. *Demography, 10*(2), 277–287.

Rogers, A. (1975). *Introduction of multiregional mathematical demography*. New York: Wiley.

Rogers, A. (1978a). Model migration schedules: An application using data for the Soviet Union. *Canadian Studies in Population, 5*, 85–98.

Rogers, A. (1978b). Special IIASA issue on migration and settlement. *Environment and Planning A, 10*(5), 469–474.

Rogers, A. (1985). *Regional population projection models*. Beverly Hills: Sage Publications.

Rogers, A. (1986). Parameterized multistate population-dynamics and projections. *Journal of the American Statistical Association, 81*(393), 48–61.

Rogers, A. (1988). Age patterns of elderly migration: An international comparison. *Demography, 25*(3), 355–370.

Rogers, A. (1995). *Multiregional demography: Principles, methods, and extensions*. Chichester: Wiley.

Rogers, A., (Ed.). (1999). Special issue: The indirect estimation of migration. *Mathematical Population Studies, 7*(3), 181–216.

Rogers, A., & Castro, L. J. (1981). *Model migration schedules*. Research report 81–30. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A., & Castro, L. J. (1986). Migration. In A. Rogers & F. Willekens (Eds.), *Migration and settlement: A multiregional comparative study* (pp. 157–208). Dordrecht: D. Reidel.

Rogers, A., Castro, L. J., & Lea, M. (2005). Model migration schedules: Three alternative linear parameter estimation methods. *Mathematical Population Studies, 12*(1), 17–38.

Rogers, A., Jones, B., Partida, V., & Muhidin, S. (2007). Inferring migration flows from the migration propensities of infants: Mexico and Indonesia. *Annals of Regional Science, 41*(2), 443–465.

Rogers, A., & Jordan, L. (2004). Estimating migration flows from birthplace-specific population stocks of infants. *Geographical Analysis, 36*(1), 38–53.

Rogers, A., & Little, J. S. (1994). Parameterizing age patterns of demographic rates with the multiexponential model schedule. *Mathematical Population Studies, 4*(3), 175–195.

Rogers, A., & Liu, J. (2005). Estimating directional migration flows from age-specific net migration data. *Review of Urban and Regional Development, 17*(3), 177–196.

Rogers, A., & Raymer, J. (1998). The spatial focus of U.S. interstate migration flows. *International Journal of Population Geography, 4*(1), 63–80.

Rogers, A., & Raymer, J. (1999). Fitting observed demographic rates with the multiexponential model schedule: An assessment of two estimation programs. *Review of Urban and Regional studies, 11*(1), 1–10.

Rogers, A., & Raymer, J. (2005). Origin dependence, secondary migration, and the indirect estimation of migration flows from population stocks. *Journal of Population and Research, 22*(1), 1–19.

Rogers, A., Raymer, J., & Newbold, K. B. (2003). Reconciling and translating migration data collected over time intervals of differing widths. *Annals of Regional Science, 37*(4), 581–601.

Rogers, A., & Sweeney, S. (1998). Measuring the spatial focus of migration patterns. *The Professional Geographer, 50*(2), 232–242.

Rogers, A., & Watkins, J. (1987). General versus elderly interstate migration and population redistribution in the United States. *Research on Aging, 9*(4), 483–529.

Rogers, A., Watkins, J. F., & Woodward, J. (1990). Interregional elderly migration and population redistribution in four industrialized countries: A comparative analysis. *Research on Aging, 12*(3), 251–293.

Rogers, A., & Willekens, F. (1986). *Migration and settlement: A multiregional comparative study*. Dordrecht: Reidel.

Rogers, A., Willekens, F., Little, J., & Raymer, J. (2002). Describing migration spatial structure. *Papers in Regional Science, 81*(1), 29–48.

Rogers, A., Willekens, F., & Raymer, J. (2001). Modeling interregional migration flows: Continuity and change. *Mathematical Population Studies, 9*, 231–263.

Rogers, A., Willekens, F., & Raymer, J. (2002). Capturing the age and spatial structures of migration. *Environment and Planning A, 34*, 341–359.

Rogers, A., Willekens, F., & Raymer, J. (2003). Imposing age and spatial structures on inadequate migration-flow datasets. *Professional Geographer, 55*(1), 56–69.

Rogerson, P. A. (1990). Migration analysis using data with time intervals of differing widths. *Papers of the Regional Science Association, 68*, 97–106.

Rowland, D. T. (2003). *Demographic methods and concepts.* Oxford: Oxford University Press.

Schmertmann, C. P. (1992). Estimation of historical migration rates from a single census: Interregional migration in Brazil 1900–1980. *Population Studies, 46*(1), 103–120.

Shryock, H. S. J. (1964). *Population mobility within the United States*. Chicago: Community and Family Study Center, University of Chicago.

Siegel, J. S., & Swanson, D. A. (Eds.). (2004). *The methods and materials of demography.* Amsterdam: Elsevier.

Smith, P. W. F., Raymer, J., & Giulietti, C. (2010). Combining available migration data in England to study economic activity flows over time. *Journal of the Royal Statistical Society. Series A, 173*(4).

Snickars, F., & Weibull, J. W. (1977). A minimum information principle: Theory and practice. *Regional Science and Urban Economics, 7*(1–2), 137–168.

SPSS. (2004). *SPSS for Windows, Release 13.0*. Chicago, IL: SPSS, Inc.

Stillwell, J. C. H. (1986). The analysis and projection of interregional migration in the United Kingdom. In R. Wood & P. Rees (Eds.), *Population structures and models: Developments in spatial demography* (pp. 160–292). London: Allen and Unwin.

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

U.S. Census Bureau. (1987). *Geographic mobility: 1985*. Current Population Reports, P-20, No. 420. Washington, DC: U.S. Government Printing Office.

U.S. Census Bureau. (2008). *A compass for understanding and using American Community Survey data: What general data users need to know*. Washington, DC: U.S. Government Printing Office.

U.S. Census Bureau. (2009). *A compass for understanding and using American Community Survey data: What PUMS data users need to know*. Washington, DC: U.S. Government Printing Office.

United Nations. (1967). *Manual IV: Methods of estimating basic demographic measures from incomplete data*. New York: United Nations: Department of Economic and Social Affairs.

United Nations. (1983). *Manual X: Indirect techniques for demographic estimation*. New York: United Nations: Department of International Economic and Social Affairs.

United Nations. (1992). *Preparing migration data for subnational population projections.* New York, NY: Department of International Economic and Social Affairs, United Nations.

Watkins, J. F. (1984). *A generalized linear program for linear parameter estimation in model migration schedules.* (Working paper 84-4). Boulder, CO: Population Program, Institute of Behavioral Science, University of Colorado.

Warren, R., & Kraly, E. P. (1985). *The elusive exodus: Emigration from the United States* (Population Trends and Public Policy Occasional Paper, No. 8). Washington, DC: Population Reference Bureau.

Warren, R., & Peck, J. M. (1980). Foreign-born emigration from the United States: 1960–1970. *Demography, 17*, 71–84.

Wetrogan, S., & Long, J. (1990). Creating annual state-to-state migration flows with demographic detail. In United States. Bureau of the Census (Ed.), *Perspectives on migration analysis*. Current population reports, P-23, No. 166. Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Willekens, F. (1982a). Multidimensional population analysis with incomplete data. In K. C. Land & A. Rogers (Eds.), *Multidimensional mathematical demography* (pp. 43–111). New York: Academic.

Willekens, F. (1982a). Specifications and calibrations of spatial interaction models: A contingency table perspective and an application to intra-urban migration in Rotterdam. *Journal of Economic and Social Geography, 74*, 239–252.

Willekens, F. (1983a). Log-linear modeling of spatial interaction. *Papers of the Regional Science Association, 52*, 187–205.

Willekens, F. (1983b). Specification and calibration of spatial interaction models: A contingency-table perspective and an application to intra-urban migration in Rotterdam. *Tijdschrift voor Economische en Sociale Geografie, 74*(4), 239–252.

Willekens, F. (1994). Monitoring international migration in Europe. *European Journal of Population, 10*(1), 1–42.

Willekens, F. (1999). Modeling approaches to the indirect estimation of migration flows: From entropy to EM. *Mathematical Population Studies, 7*, 239–278.

Willekens, F., & Baydar, N. (1986). Forecasting place-to-place migration with generalized linear models. In R. I. Wood & P. Rees (Eds.), *Population structures and models: Developments in spatial demography* (pp. 203–244). London: Allen and Unwin.

Willekens, F., Por, A., & Raquillet, R. (1981). Entropy, multiproportional, and quadratic techniques for inferring patterns of migration from aggregate data. In A. Rogers (Ed.), *Advances in multiregional demography*. Research report 81–6. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Willekens, F., & Rogers, A. (1978). *Spatial population analysis: Methods and computer programs*. Research report 78–18. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Willekens, F. (1980). Entropy, multiproportional adjustment and the analysis of contingency tables. *Systemi Urbani, 2*, 171–201.

Wilson, A. G. (1970). *Entropy in urban and regional modelling*. London: Pion.

Yano, K. (1991). The integration of spatial interaction models using generalized linear modeling. *Geographical Review of Japan, 64*(6), 367–387.

Zaba, B. (1987). The indirect estimation of migration: A critical review. *International Migration Review, 21*, 1395–1445.

Index