Imposing Age and Spatial Structures on Inadequate Migration-Flow Datasets*

Andrei Rogers
University of Colorado

Frans Willekens
University of Groningen, Netherlands

James Raymer
University of Colorado

With the elimination of the long-form questionnaire from future decennial censuses and its replacement by a much smaller continuous monthly sampling survey (the American Community Survey), students of territorial mobility may find it necessary to deal with inadequate, missing, or inaccurate sample data on migration by adopting an approach that "improves" such data using information from different geographical areas, time periods, and data sources. We develop such an approach in this article and illustrate it with interregional migration flow data reported by the U.S. decennial censuses of 1980 and 1990 and by the 1985 Current Population Survey. **Key Words: estimation, migration, models.**

Introduction

Under pressure from Congress to reduce the cost of the national decadal census, the U.S. Census Bureau plans to replace its long form in the year 2010 census with a continuous monthly sampling survey based on its successful Current Population Survey (CPS). Called the American Community Survey (ACS), the data this survey would provide would have the advantages of lower cost and more up-to-date information, but such a replacement would greatly complicate the measurement and analysis of internal migration flows for use in subnational projections, for example. Even though the current sample size of the CPS is to be increased significantly, it will still be several orders of magnitude too small to permit analyses at currently conventional levels of geographical detail—counties, metropolitan areas, even states. The remedy proposed here is the imposition of age and spatial structures on observed migration-flow data using migration data drawn from other sources—in particular, past data.

Traditionally, the methods developed by demographers to impose empirical regularities captured in other datasets adopt stylized profiles of age-specific rates or proportions, known as model schedules. When data for a given population are incomplete, a schedule is selected for a comparable population other than the population being studied and then modified to fit. This is called indirect estimation. The justification for such an approach is that age profiles of an observed schedule 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.

In general, model schedules take one of two distinct approaches to summarizing a demographic set of rates in terms of relatively few parameters: functional (analytical) representations and relational representations. Functional representations describe the observed age pattern of an entire schedule by a mathematical curve. The Heligman-Pollard (1980) model mortality schedule and the Rogers-Castro (1981) model migration schedule are examples of this class of model schedules. Relational representations, on the other hand, describe an observed age pattern by associating it with a "standard" pattern and capturing its deviations from that pattern with a few parameters. The

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

The Professional Geographer, 55(1) 2003, pages 56–69 © Copyright 2003 by Association of American Geographers.
Initial submission, August 2001; revised submission, May 2002; final acceptance, June 2002.
Published by Blackwell Publishing, 350 Main Street, Malden, MA 02148, and 9600 Garsington Road, Oxford, OX4 2DQ, UK.

Brass (1974) model mortality schedule and the Coale-Trussell (1974) model fertility schedule are members of this class of model schedules.

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, the United Nations manual on the subject (United Nations 1983, 1) is very explicit in its noncoverage 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." More recently, a chapter on indirect estimation methods in an important text on formal demography (Preston, Heuveline, and Guillot 2001) totally ignores migration.

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 such "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 must also focus on spatial patterns. This is where the geographer's particular contribution to migration analysis becomes evident. For decades, geographers have focused on models for describing the spatial patterns of migration. Early efforts using gravity models (Isard 1960) gave way to entropy maximizing models (Wilson 1970), which in turn gave way to log-linear formulations of very general spatial interaction models (Willekens 1983; Sen and Smith 1995).

Log-linear models have a long history. The theory was developed in the sixties and seventies to model count data that are categorical. Bishop, Fienberg, and Holland (1975) provide a basic text. When data are counts or frequencies, they may be viewed as being generated by a Poisson process. The number of migrations by migrant category is an example of count data. Since a count or frequency is necessarily nonnegative, a logarithmic transformation is used to assure that any prediction of the count is nonnegative. When the logarithmic transformation of the count is a linear function of a set of parameters, the model is a log-linear model. The number of parameters that can be included in the model depends on data availability. When data are complete, a model can be specified that includes a parameter for each observation. That model is a saturated model. When data are incomplete, some parameters cannot be estimated from the data. The best alternative is to try to obtain parameter estimates from other relevant information, such as a historical pattern that is considered sufficiently close to the current but unobserved pattern. The historical interaction pattern can be imposed onto the current migration pattern using, for example, iterative proportional fitting (IPF).

Geographers have used IPF methods to, for example, "adjust a two-dimensional matrix iteratively until the row sums and column sums equal some predefined values" (Wong 1992, 340; Johnston and Pattie 1993). And they have recognized the value of adopting the method of offsets as a means for imposing historical spatial interaction patterns on a current flow matrix (Knudsen 1992). IPF methods have a long history. Bartlett (1935) is generally believed to have been the first to describe a method of getting maximum likelihood estimates (MLEs) for a model that does not possess closed-form solutions (see Fienberg 1970; Bishop, Fienberg, and Holland 1975, 83). The algorithm itself is often associated with Deming and Stephan (1940). Instead of the maximum-likelihood criterion, some authors use a criterion that maximizes the entropy (see, e.g., Wilson 1970) or minimizes the information gain subject to constraints (see, e.g., Gokhale and Kullback 1978; Haynes and Phillips 1982). The different perspectives on the estimation problem have been the subject of extensive discussion in the geographic literature in the 1980s and beyond (see, e.g., Tyree 1973; Plane 1981; Flowerdew and Aitkin 1982; Snickars and Weibull 1977; Willekens, Por, and Raquillet 1979; Flowerdew and Lovett 1988; Fotheringham and O'Kelly 1989; Willekens 1994). We draw on such procedures to develop a demographically sound method for imposing age and spatial structures on inadequate migration flow datasets. Such an application contributes to the growing literature on the "indirect estimation" of migration. See, for example, the special issue of *Mathematical Population Studies* devoted to this topic (Rogers 1999).

The imposition of observed regularities in both the age and spatial patterns of interregional migration to "discipline" inadequate data on territorial mobility holds great promise as a means for developing improved age- and destination-specific migration-flow data from inaccurate and partial information on this most fundamental process underlying population redistribution. In this article, we present a method that adopts a relational perspective. The age and spatial patterns of migration are related, not to a standard, but to historical patterns of migration. The historical patterns, and assumptions regarding trends, are used as a basis for improving observed migration-flow data. However, such preliminary "predictions" could also involve a standard. Indeed, the (re)construction of migration flows may involve the combination of information from several data sources. The main feature of the proposed method is that we use a loglinear model to capture the contribution 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 the various datasets.

We begin the article by analyzing log-linear models of interregional migration and examine migration data from two sources: the U.S. Census (1984, 1993) and the CPS (U.S. Census Bureau 1987). The CPS is used because the U.S. Census Bureau's forthcoming ACS will resemble the CPS. The decennial census provides data for the periods between 1975 and 1980 and between 1985 and 1990; the CPS provides data for the period between 1980 and 1985. Next, we address the problem of the interpretation of the parameters of the model, a source of possible confusion. 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. Our consideration of log-linear models includes the conventional log-linear model and also the log-linear model with an offset whereby some of the parameters are fixed exogenously (Lin 1999a, 1999b). The indirect estimation of migration is considered next. The main issues addressed in that section include (1) how to use log-linear models to predict migration flows, and (2) how to derive adequate values for the parameters from the available information on migration. The information considered consists of migration statistics (quantitative data) and judgmental (qualitative) data. Finally, we end the article with a brief discussion and a few conclusions.

Log-Linear Models of the Structures of Interregional Migration Flows

The method proposed in this article for the improved or indirect estimation of migration flows makes extensive use of the parameters of log-linear models and their interpretation. The interpretation of such parameters relies heavily on the concept of odds: the ratio of two frequencies (counts). In the case of a binary variable, the odds are the ratio of the frequency of falling into one category over the frequency of not falling into that category. In the case of a polytomous variable, the odds are the ratio of the frequency in a given category over the frequency in the reference category. Odds are a way of expressing the frequency of an event that is consistent with the associated log-linear model.

The log-linear model is used in this section to describe migration flows. The migration data considered are interregional migrations in the United States for three periods: 1975–1980, 1980–1985, and 1985–1990. The data for the 1975–1980 and 1985–1990 periods come from the U.S. Public Use Microdata Series files. The 1980–1985 migration data come from a CPS report (U.S. Census Bureau 1987). 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 Census Bureau. The 1975–1980 and 1985–1990 migration data are based on a much larger sample size (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 will be viewed with some question.

Modeling Origin-Destination Migration Flows

Table 1 sets out the migration-flow tables. The log-linear model that describes the data

Table 1 *U.S. Interregional Migration Flows (in Thousands): 1975–1980, 1980–1985, and 1985–1990*

		Region of Destination				
Period	**Region of Origin**	**Northeast**	**Midwest**	**South**	**West**	**Total**
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

perfectly is the *saturated* log-linear model. Table 2 sets out the parameters for the saturated log-linear model estimated separately for the 1975–1980, 1980–1985, and 1985–1990 migration-flow tables in Table 1. Note that the last category, the West region, is the reference category. The multiplicative log-linear model that produced those parameters is specified as:

$$\hat{n}_{ij} = \tau\tau_i^O\tau_j^D\tau_{ij}^{OD} \qquad (1)$$

where $\hat{n}_{ij}$ denotes the predicted number of migrants from region i to region j and the τs denote the parameters of the model, consisting of an overall effect (or intercept), τ, an origin main effect, τ_i^O, a destination main effect, τ_j^D, and an origin-destination interaction effect, τ_{ij}^{OD}.

The interpretation of the parameters of log-linear models relies heavily on the concept of odds. If the West region is the reference category, the effects on migration flows of origins, destinations, and their interactions are expressed relative to the effect that the West region has on those migration flows. The overall effect of the saturated model is the number of stayers in the West (37,902 thousand in the period between 1975 and 1980). The main effect associated with a given origin is the odds that a migrant to the West comes from that origin rather than from the West. For instance, τ_1^O is the odds that a migrant to the West comes from the Northeast rather than from the West. For the period between 1975 and 1980, it is equal to 753/37,902 = 0.0199.

Table 2 *Saturated Log-Linear Model Parameters of U.S. Interregional Migration Flows: 1975–1980, 1980–1985, and 1985–1990*

Parameter	**1975–1980**	**1980–1985**	**1985–1990**
τ	37,902	40,028	43,733
τ_1^O	0.0199	0.0118	0.0124
τ_2^O	0.0335	0.0286	0.0234
τ_3^O	0.0301	0.0256	0.0289
τ_1^D	0.0076	0.0059	0.0089
τ_2^D	0.0179	0.0167	0.0161
τ_3^D	0.0295	0.0271	0.0269
τ_{11}^{OD}	7,562.8244	16,012.1006	9,222.4868
τ_{12}^{OD}	34.3499	47.9424	40.9356
τ_{13}^{OD}	80.8989	108.1804	125.0358
τ_{21}^{OD}	36.4230	48.1297	41.4588
τ_{22}^{OD}	2,255.8903	2,735.8570	3,165.2901
τ_{23}^{OD}	49.2003	63.0104	63.9627
τ_{31}^{OD}	80.4391	107.3721	75.5731
τ_{32}^{OD}	53.0906	49.9589	61.0016
τ_{33}^{OD}	1,990.0201	2,476.4962	2,142.4386

Notes: The superscripts O and D equal region of origin and region of destination, respectively. The subscripts 1, 2, 3, and 4 equal the regions of the Northeast, Midwest, South, and West, where the West region is the reference category.

The main effect associated with a given destination is the odds that a migrant from the West selects that particular region as a destination, rather than the West. For instance, τ_1^D is the odds that an out-migrant from the West settles in the Northeast rather than in the West. For the period between 1975 and 1980, it is equal to 287/37,902 = 0.0076. The interaction effects are ratios of odds. For instance, τ_{12}^{OD} is the odds that an out-migrant from the

Northeast selects the Midwest rather than the West, divided by the odds that an out-migrant from the West selects the Midwest rather than the West. For 1975–1980, it is [462/753]/[677/37,902] = 34.3499.

Modeling Origin-Destination Migration Flows with Prior Information

If the data are incomplete, auxiliary information may be used to predict migration flows. Let n^*_{ij} denote a historical (or hypothetical) migration-flow table. The migration-flow table for the current period may be predicted on the basis of, for example, information regarding the *aggregate* total number of persons living in regions i and j at the beginning and end of the time interval, n_{i+} and n_{+j}, respectively, and the historical data on the number of origin-destination-specific migration flows represented by n^*_{ij}. The model, then, is:

$$\hat{\theta}_{ij} = \frac{\hat{n}_{ij}}{n^*_{ij}} = \frac{\tau\tau_i^O\tau_j^D\tau_{ij}^{OD^*}}{\tau^*\tau_i^{O^*}\tau_j^{D^*}\tau_{ij}^{OD^*}} = \frac{\tau\tau_i^O\tau_j^D}{\tau^*\tau_i^{O^*}\tau_j^{D^*}}$$
$$= \nu\nu_i^O\nu_j^D \qquad (2)$$

where $\hat{\theta}_{ij}$ is the odds between the predicted number of migrants or stayers, $\hat{n}_{ij}$, and the corresponding number included in the offset, n^*_{ij}. Note that $\hat{n}_{ij} = n^*_{ij}\hat{\theta}_{ij} = n^*_{ij}\nu\nu_i^O\nu_j^D$. The *ν*s denote the parameters of the log-linear-with-offset model. The parameters of this model are related to the saturated log-linear models discussed above (i.e., equation [1]), with $\nu = \tau/\tau^*$, $\nu_i^O = \tau_i^O/\tau_i^{O^*}$, and $\nu_j^D = \tau_j^D/\tau_j^{D^*}$ with the numerator being equal to the saturated log-linear parameters of the predicted number of migrants or stayers and the denominator being equal to the saturated log-linear model parameters of the offset. The model in equation (2) is not a saturated model and therefore borrows the origin-destination interaction effect parameters, τ_{ij}^{OD}, from the auxiliary (e.g., historical) data. Notice that the τ_{ij}^{OD} parameter in the numerator and denominator have asterisks; this implies that they come from the offset and that they are equal to each other (thus, they cancel each other out). The result of the above model is a migration-flow table that exhibits the level of a current period but adopts the spatial structure of the offset (for example, of the historical pattern).

To illustrate the method, 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. Table 3 sets out the resulting predicted migration-flow table and ratios of predicted-to-observed migration-flow tables for the period between 1980 and 1985. The parameters of the log-linear-with-offset model are set out in Table 4, along with the saturated log-linear parameters of the predicted values ($\hat{n}_{ij}$). Note that the interaction parameters of the saturated log-linear model are equal to those set out in Table 2 for the period between 1975 and 1980. For example, the number of migrants from the Northeast to the South between 1980 and 1985 predicted by the model is

$$\hat{n}_{13} = n^*_{13}\nu\nu_1^O\nu_3^D$$
$$= (1,800)(1.0469)(0.8014)(1.0687)$$
$$= 1,614$$

These parameters also can be obtained by dividing the saturated log-linear parameters of the *predicted* 1980–1985 migration-flow matrix (in Table 4) by the saturated log-linear parameters of the *observed* 1975–1980 migration-flow matrix (in Table 2): $\nu = 39{,}678/37{,}902 = 1.0469$, $\nu_1^O = 0.0159/0.0199 = 0.8014$, and, $\nu_3^D = 0.0316/0.0295 = 1.0687$.

In the conventional log-linear model, the main effects are odds. In the log-linear model with an offset, the main effects are odds ratios. The origin main effect, $\nu_1^O = 0.8014$, denotes the ratio of the odds that a migrant from the Northeast goes to the West relative to stayers in the West divided by the corresponding odds in the reference period. The odds predicted by the model (Table 3) are 632/39,678 = 0.0159 and the odds in the reference period (Table 1) are 753/37,902 = 0.0199. The corresponding ratio is 0.8014. The effect indicates the change in the odds that a migrant comes from the Northeast. The odds declined 20% from 1975–1980 to 1980–1985 (if the Census data and the CPS data are comparable). Note that the odds ratio is the same for all regions of origin. For instance, the ratio of the predicted odds that person in the Northeast comes from the Northeast rather than from the West over the odds in the reference period is [44,445/369]/[43,123/287] = 0.8014.

The main effect of the destination, $\nu_3^D = 1.0687$, is also an odds ratio. It is the odds that a

Table 3 *Predicted Migration-Flow Table and Ratios of Predicted to Observed Migration Flows for the 1980–1985 Period for the Log-Linear Model with the 1975–1980 Migration-Flow Table as the Offset*

	Region of Destination				
Region of Origin	**Northeast**	**Midwest**	**South**	**West**	**Total**
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
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

Table 4 *Parameters Used to Predict the Migration-Flow Table for the 1980–1985 Period Using the 1975–1980 Period as the Offset*

Log-Linear with Offset Parameters		**Saturated Log-Linear Parameters of the Predicted Values**	
ν	1.0469	τ	39,678
ν_1^O	0.8014	τ_1^O	0.0159
ν_2^O	0.9579	τ_2^O	0.0321
ν_3^O	0.9103	τ_3^O	0.0274
ν_1^D	1.2285	τ_1^D	0.0093
ν_2^D	1.0152	τ_2^D	0.0181
ν_3^D	1.0687	τ_3^D	0.0316
		τ_{11}^{OD}	7,562.8244
		τ_{12}^{OD}	34.3499
		τ_{13}^{OD}	80.8989
		τ_{21}^{OD}	36.4230
		τ_{22}^{OD}	2,255.8903
		τ_{23}^{OD}	49.2003
		τ_{31}^{OD}	80.4391
		τ_{32}^{OD}	53.0906
		τ_{33}^{OD}	1,990.0201

Notes: The ν-parameters denote the parameters of the log-linear with offset model. The τ-parameters denote the parameters of the log-linear model. The superscripts O and D equal region of origin and region of destination, respectively. The subscripts 1, 2, 3, and 4 equal the regions of the Northeast, Midwest, South, and West, where the West region is the reference category.

resident of the West migrates to the South rather than staying in the West (predicted by the model) divided by the corresponding odds in the reference period. The odds predicted by the model is 1,253/39,678 = 0.0316 and the odds in the reference period is 1,120/37,902 = 0.0295. Thus the ratio is 1.0687.

The origin-destination interaction effects exhibited by the matrix of predicted migration flows in 1980–1985 are identical to the effects exhibited by migration in the reference period (1975–1980). For example, consider the migration from the Northeast to the South. The interaction effect is the odds that an out-migrant from the Northeast selects the South rather than the West, divided by the odds that an out-migrant from the West goes to the South rather than to the West. For 1975–1980, it is [1,800/753]/[1,120/37,902] = 80.8989. For the migration in 1980–1985, predicted by the model, it is [1,614/632]/[1,253/39,678] = 80.8989. The prediction of migration with offsets preserves the interaction effects, expressed as relative odds. The odds that a resident of the Northeast selects the South relative to the odds of a resident of the West is being preserved during the prediction process. This unambiguous interpretation of the parameters of the model is a particularly interesting feature of the log-linear model, which is not shared by many other relational models. It is that feature that will be used extensively in the indirect estimation of migration. But before embarking on the task of indirect estimation, we first extend the log-linear model to age-specific migration flows.

Modeling Age-Specific Origin-Destination Migration Flows

Only a dozen five-year age groups are distinguished in our analysis, ranging from the 0–4 age group to the 55–59 age group. The reason is that the published CPS on interregional migration does not provide age detail beyond age 60 (the age in 1980). Moreover, the published CPS data on migration are for five-year

age groups up to age 34 and for ten-year age groups for ages 35 and higher. The ten-year age data were disaggregated into five-year data by assuming a uniform distribution of migrants in the ten-year age interval, an assumption that is considered to be adequate for the illustration in this article.

The log-linear model associates a parameter with each age category. Thus the twelve age groups require twelve age parameters. The last age group (55–59) is adopted as the reference category. In addition, parameters are associated with the various possible interactions between origin, destination, and age. Consider the saturated model of migration by origin, destination, and age:

$$\hat{n}_{ij}(x) = \tau\tau_i^O\tau_j^D\tau^A(x)\tau_{ij}^{OD}\tau_i^{OA}(x)\tau_j^{DA}(x)\tau_{ij}^{ODA}(x) \quad (3)$$

where the superscript A denotes an age effect and x denotes the age-group category measured at the beginning of the migration time interval. Estimating the model's parameters with data given for 1975 to 1980 in thousands and age in years in 1975 produces the results illustrated in Figure 1.

For illustration, consider the Northeast to South flow (in thousands) for persons aged 20 to 24 during the period between 1975 and 1980:

$$\begin{aligned}\hat{n}_{13}(20) &= \tau\tau_1^O\tau_3^D\tau^A(20)\tau_{13}^{OD}\tau_1^{OA}(20)\tau_3^{DA}(20) \\ &\quad \times \tau_{13}^{ODA}(20) \\ &= (1,769.9384)(0.0111)(0.0130) \\ &\quad \times (2.0799)(380.4672)(3.3488) \\ &\quad \times (3.7547)(0.0886) \\ &= 225\end{aligned}$$

The overall effect, $\tau = 1{,}770$, denotes the number of persons (in thousands) aged 55 to 59 who stayed in the West during the period between 1975 and 1980. The main effects of origin ($\tau_1^O = 0.0111$), destination ($\tau_3^D = 0.0111$), and age ($\tau^A(20) = 0.0111$) are odds interpreted as, respectively the ratio of persons aged 55 to 59 migrating from the Northeast to the West relative to stayers in the West, the ratio of persons aged 55 to 59 from the West to the South relative to stayers in the West, and the ratio of persons aged 20 to 24 staying in the West relative to persons aged 55 to 59 staying in the West.

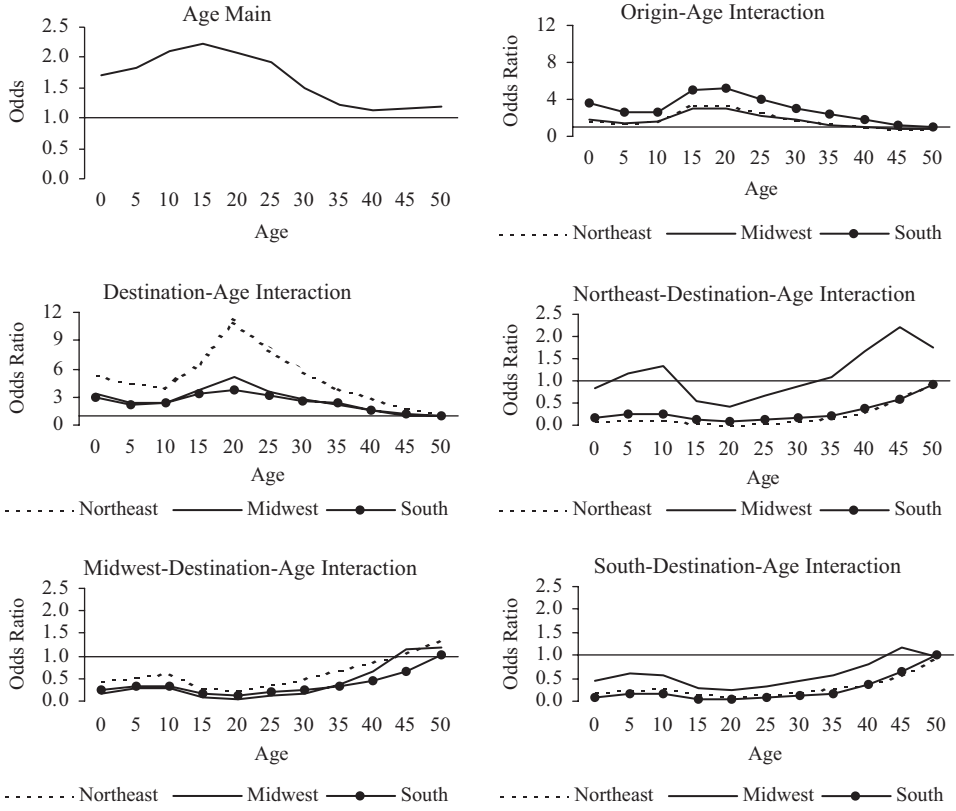
The origin-destination interaction effect ($\tau_{13}^{OD} = 380.4672$) denotes, for persons aged 55 to 59, the odds ratio of (1) the odds of migrating from the Northeast to the South relative to the odds of migrating from the Northeast to the West to (2) the odds of migrating from the West to the South relative to stayers in the West. The origin-age interaction effect ($\tau_1^{OA}(20) = 3.3488$) denotes the odds ratio that a person who migrates from the Northeast to the West is 20 to 24 rather than 55 to 59, divided by the corresponding odds for a stayer in the West. The destination-age interaction effect ($\tau_3^{DA}(20) = 3.7547$) denotes the odds ratio that a person who migrates from the West to the South is 20 to 24 rather than 55 to 59, divided by the corresponding odds for a stayer in the West. Finally, the origin-destination-age interaction effect ($\tau_{13}^{ODA}(20) = 0.0886$) is an odds ratio that represents (1) the odds ratio of migrating from the Northeast to the South relative to the West at age 20 to 24 to the corresponding odds at age 55 to 59 to (2) the odds ratio of migrating from the West to the South relative to staying in the West at age 20 to 24 to the corresponding odds at age 55 to 59.

When we introduce an offset, the values of some of the parameters are "borrowed" from the offset. Which parameter values are taken from the offset depends on the specification of the model—more specifically, on the effect parameters that are included. For example, consider the main effects model:

$$\hat{n}_{ij}(x) = n_{ij}^*(x)\nu\nu_i^O\nu_j^D\nu^A(x) \quad (4)$$

and assume that it is estimated to predict the migration flows between 1980 and 1985. The data required for this model consist of the offset, which describes the origin-destination migration flows by age for the period between 1975 and 1980, the arrivals and departures by region during the period between 1980 and 1985, and the main effects age structure of the 1980–1985 migrants. The two-way and three-way interaction effects are borrowed from the 1975–1980 migration data.

The shortcomings of this approach are twofold. First, the number of parameters to be added is large, since a parameter is associated with each age category. Second, the approach does not assure that the age profiles of



Overall effect, origin main effect, destination main effect, and origin-destination interaction effect parameters:

τ 1,769.9384	τ_1^0 0.0111	τ_1^D 0.0016	τ_{11}^{OD} 75,856.6004	τ_{21}^{OD} 74.2249	τ_{31}^{OD} 343.6075
	τ_2^0 0.0197	τ_2^D 0.0070	τ_{12}^{OD} 38.9508	τ_{22}^{OD} 10,306.1904	τ_{32}^{OD} 107.7162
	τ_3^0 0.0106	τ_3^D 0.0130	τ_{13}^{OD} 380.4672	τ_{23}^{OD} 173.1572	τ_{33}^{OD} 13,522.3615

Figure 1 *Saturated log-linear parameters for the origin-destination-age model.*

migration patterns follow well-established regularities. In the next section, therefore, we impose observed regularities in the age profile of migration. This involves the parameterization of the pattern of the twelve age groups using model migration schedules (Rogers and Castro 1981). The parameterization assures a well-established profile, while reducing the number of main effect (age effect) parameters from twelve to seven.

Imposing Structures on Observed Migration Flows

The log-linear models specified in the section above are applied in this section to demonstrate how the relational method can be used to impose structure on observed interregional migration flows. Several types of information on migration might be available to do this. The information may be quantitative or qualitative. Quantitative information typically consists of historical and/or contemporary migration statistics. Qualitative data might consist of judgments about migration patterns and changes in patterns. Such judgmental data are often used in forecasting but have not been used as much in the estimation (prediction) of migration. In this section, we consider quantitative data first. Then we consider how judgmental data might be introduced.

Aggregate Migration Flows

Historical information from the 1980 census was combined with information from the 1985 CPS in the illustration set out in the previous section. The information consisted of the population by region in 1985 as revealed by the CPS of 1985 and the distribution of the total number of survivors in 1985 by region of residence in 1980. These data constituted the marginal totals of the 1980–1985 migration matrix. Equation (2) presented the associated log-linear model. Its parameters denoted the results of the prediction relative to the situation in the reference period (1975–1980) and described (1) the predicted growth in the number of persons in the reference category (i.e., stayers in the West), (2) the growth in the odds that migrants into any of the regions came from a given region rather than from the reference region (West), and (3) the growth in the odds that migrants from any region selected a given region rather than the reference region. For instance, the number of stayers in the West was predicted to grow by 4.7%, from 37,902 to 39,678. The odds that an in-migrant to the West came from the Northeast rather than from the West was predicted to decline by about 20% from 0.0199 to 0.0159 (i.e., the odds predicted for the period between 1980 and 1985 were 80.14% of the odds between 1975 and 1980). But the observed decline was considerably larger: 60%, a drop from 0.0199 to 0.0118. The conversion of this change in odds into a corresponding change in the proportion of in-migrants in the West coming from the Northeast is straightforward. It is equal to $[\hat{\theta}_1/(\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 + \hat{\theta}_4)]/[\theta_1^*/(\theta_1^* + \theta_2^* + \theta_3^* + \theta_4^*)]$, where $\hat{\theta}_i$, refers to the odds predicted by the model and $*$ refers to the reference period—the 1975–1980 period. Note that $\hat{\theta}_1/\theta_1^*$ is the main effect parameter associated with the origin Northeast (0.8017). A decline in the odds that a resident in the West in 1985 is an in-migrant from the Northeast of 20%, from 0.0199 to 0.0159, implies a decline of 19% (i.e., from 1.83% to 1.48%) in the proportion of recent in-migrants from the Northeast.

The above analysis assumes that the contemporary data consist of arrivals and departures by region (including stayers). In practice, however, other types of data may be available instead. Suppose that a migration survey was carried out in only one region (e.g., the West) and the numbers of arrivals and departures are available for that region only—that is, the data are shown in the fourth column and the fourth row of the 1980–1985 migration matrix in Table 1. From that limited data, we may derive the parameters of the log-linear model by using the 1975–1980 migration matrix as an offset. The interpretation of the parameters of the log-linear model with that particular offset implies that the offset and the *predicted* values of arrivals and departures of the reference region suffice to predict the entire migration-flow matrix.

Let the period between 1975 and 1980 be the reference period determining the offset, and assume that the migrant-flow matrix for the period between 1980 and 1985 needs to be predicted. Table 5A shows the predicted migration flows. The parameters can be derived from the information provided in Tables 1 (1975–1980 migration-flow matrix) and 5A. The overall effect parameter (ν) is the ratio of stayers in the West between 1980 and 1985 predicted by the model and the number of stayers between 1975 and 1980. It is 40,028/37,902 = 1.0561. The origin main effect parameters are:

$$\nu_1^O = [473/40,028]/[753/37,902] = 0.5948$$

$$\nu_2^O = [1,144/40,028]/[1,269/37,902] = 0.8536$$

$$\nu_3^O = [1,024/40,028]/[1,141/37,902] = 0.8498$$

The destination main effect parameters are:

$$\nu_1^D = [237/40,028]/[287/37,902] = 0.7819$$

$$\nu_2^D = [669/40,028]/[677/37,902] = 0.9357$$

$$\nu_3^D = [1,085/40,028]/[1,120/37,902] = 0.9173$$

The decline of the West as a destination in the early 1980s is reflected in the other regions. As a consequence, the overall level of migration is underestimated (169,000 instead of the observed 216,000). The reference region experienced a change in migration that was considerably different from that of the other regions—hence the inaccurate predictions of

Table 5 *Predicted Migration Flows for the 1980–1985 Period Using Data on the West Region Only*

	Region of Destination				
Region of Origin	**Northeast**	**Midwest**	**South**	**West**	**Total**
A.Data on West Region Only					
Northeast	21,181	272	1,037	473	22,963
Midwest	247	43,135	1,526	1,144	46,052
South	488	909	55,235	1,024	57,656
West	237	669	1,085	40,028	42,019
Total	22,153	44,985	58,883	42,669	168,690
B.Data on West Augmented with Judgmental Information					
Northeast	40,243	516	2,365	899	44,023
Midwest	296	51,762	2,197	1,373	55,628
South	488	909	66,282	1,024	68,703
West	237	669	1,302	40,028	42,236
Total	41,264	53,856	72,146	43,324	210,590

the numbers of interregional migrants between 1980 and 1985.

To improve the predictions, we may introduce judgmental data. Assume that our socioeconomic studies indicate that the attractiveness of the West diminished in the early 1980s and that the South became more attractive. In addition, assume that other studies showed an increased propensity to leave the Northeast and the Midwest. That information can be incorporated into the parameter values of the log-linear model—that is, into the odds. Suppose the odds that a migrant selects the South rather than the West is 20% higher than revealed by the migration pattern to the West between 1980 and 1985, and suppose that the odds that a migrant into the West originated in the Northeast (rather than in the West) is 90% higher, and that in the Midwest it is 20% higher than revealed by the patterns of arrival in the West. The parameters of the log-linear model with offset would then change accordingly. Table 5B shows the associated revised predicted migration flows for the period between 1980 and 1985. The predictions are considerably improved, except for the migration from the Northeast to the West, which is highly overpredicted.

We could improve the model still further by adding a parameter (effect) to capture the interaction between the Northeast and the West, beyond that captured by the offset. That single parameter would correct the predicted migration from the Northeast to the West. For example, were it to be $\nu_{14}^{OD} = 0.5$, then the model *with interaction effects* would become:

$$\hat{n}_{ij} = n_{ij}^{*} \nu \nu_i^{O} \nu_j^{D} \nu_{ij}^{OD} \qquad (5)$$

if, except for $i = 1$ and $j = 4$, all other interaction parameters were set equal to one. The migration from the Northeast to the West predicted by the model would then be:

$$\hat{n}_{14} = (753)(1.0561)(1.1301)(1.0000)(0.5000) = 449,$$

which is very close indeed to the observed value of 473. An interaction effect of 0.5261 would result in a perfect prediction of the migration from the Northeast to the West without affecting the other migration flows. However, the marginal totals are affected. The procedure demonstrates how both quantitative and judgmental information on migration can be included in the model by including the appropriate parameters and by "guessing" adequate values of the parameters. The model- based approach demonstrates the strength of the log-linear model as a vehicle with which to combine different types of data from different sources.

Age-Specific Migration Flows

A particularly useful technique in the context of the indirect estimation of migration is the use of model migration schedules. Their application involves a number of steps. First, model migration schedules with, in this instance, seven parameters (Rogers and Castro 1981) need to be estimated for the four regions. Table 6 shows the resulting parameter estimates. Second, the out-migration *origin-specific* proportions *predicted* by these model migration schedules need to be applied to the resident population of 1980 to obtain the *predicted* number of out-migrants. Table 7 shows

the observed and predicted out-migration proportions and numbers of out-migrants by age and region that result. Third, the *predicted* number of out-migrants needs to be allocated to regions of destination using a set of age-specific *destination* proportions. The proportions observed between 1980 and 1985 period could be used, as could any other set of appropriate proportions. For illustrative purposes, we use the proportions observed for the period between 1975 and 1980. Since the stayers are not omitted, the diagonal elements need to be set to zero. The migration matrices that result define the offset. Such a model predicts that 220,000 persons living in the Northeast aged 20–24 in 1980 will live in the South in 1985. The observed number in the CPS was 238,000. The number predicted on the basis of the model migration schedule and the destination-choice proportions of 1975–1980 was 205,000.

Note that the offset is generated by the out-migration proportion for age 20 to 24 predicted by the model migration schedule (401,000) and the destination-choice proportion for that age group observed between 1975 and 1980 (51.14% of persons of that age who left the Northeast went to the South). The procedure illustrates the power and the flexibility of the log-linear model in the context of the indirect estimation of migration. Model schedules fitted to contemporary data replaced the observed age profiles of out-migrants and were combined with historical data on destination-choice

Table 6 *Model Migration-Schedule Parameters of Observed Age-Specific Out-Migration Proportions: 1980–1985 Interregional Migration in the U.S.*

	Conditional Survivorship Proportions			
	Northeast	**Midwest**	**South**	**West**
a_1	0.0183	0.0397	0.0279	0.0471
α_1	0.1564	0.0603	0.0822	0.0820
a_2	0.1176	0.1227	0.0835	0.0954
α_2	0.1266	0.0952	0.0986	0.0786
μ_2	15.4804	14.7267	15.3776	15.4656
λ_2	0.6000	0.6000	0.5999	0.6000
a_0	0.0310	0.0292	0.0162	0.0161

Table 7 *Observed and Predicted (from Model Migration-Schedule Parameters) Age-Specific Out-Migration: 1980–1985*

	Proportions				Flows			
	Northeast	**Midwest**	**South**	**West**	**Northeast**	**Midwest**	**South**	**West**
Observed								
0	0.0491	0.0734	0.0443	0.0659	156	322	246	228
5	0.0403	0.0551	0.0346	0.0391	140	257	197	133
10	0.0343	0.0497	0.0285	0.0426	142	237	170	147
15	0.0653	0.0967	0.0492	0.0543	290	528	324	216
20	0.0967	0.1128	0.0720	0.0903	413	621	497	400
25	0.0625	0.0793	0.0492	0.0703	248	414	316	291
30	0.0477	0.0644	0.0369	0.0441	165	255	187	140
35	0.0477	0.0644	0.0369	0.0441	165	255	187	140
40	0.0342	0.0410	0.0234	0.0306	88	111	90	65
45	0.0342	0.0410	0.0234	0.0306	88	111	90	65
50	0.0314	0.0296	0.0166	0.0183	82	84	60	37
55	0.0314	0.0296	0.0166	0.0183	82	84	60	37
Total					2,058	3,278	2,423	1,896
Predicted								
0	0.0493	0.0690	0.0442	0.0632	157	302	245	219
5	0.0394	0.0586	0.0348	0.0474	137	273	198	161
10	0.0348	0.0510	0.0285	0.0369	144	243	170	127
15	0.0657	0.0965	0.0491	0.0563	292	527	323	224
20	0.0939	0.1123	0.0714	0.0878	401	618	493	389
25	0.0665	0.0841	0.0520	0.0671	264	439	334	278
30	0.0499	0.0644	0.0384	0.0506	173	254	195	160
35	0.0410	0.0519	0.0299	0.0393	142	205	152	125
40	0.0363	0.0439	0.0246	0.0318	93	119	94	67
45	0.0338	0.0387	0.0214	0.0266	86	105	82	56
50	0.0325	0.0355	0.0194	0.0232	85	101	70	46
55	0.0318	0.0333	0.0182	0.0209	83	95	66	42
Total					2,056	3,282	2,422	1,894

proportions to generate the offset, which was then used with contemporary aggregate information on migration to predict migration flows by origin, destination, and age. The log-linear model was the vehicle that integrated all of that information.

Conclusion

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 and Pollard 1980) and relational representations (Brass 1974) of age patterns have occupied a central position in such efforts at indirect estimation (Preston, Heuveline, and Guillot 2001). The indirect estimation of migration is of a more recent date, in part because the problem is more complicated. 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.

This article proposes such a method. We have outlined a very general method for imposing structure on inadequate observed migration-flow data, which may be viewed as belonging to the class of relational models. Our general approach uses a regression model to predict migration from partial data contributed by different data sources. The different explanatory variables that are commonly used in such models are replaced by different data sources. Since 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 whether 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 datasets to the predicted migration flows.

The effectiveness of the proposed method depends on the interpretation of the parameters of the log-linear model. When new data on migration are added to existing data, the contribution of the new data depends on which of the parameters are affected and how they are affected. A particularly useful feature of the proposed method is that the new data may be both quantitative and qualitative (judgmental) data. The method allows the statistical data to be combined with judgments or expert opinions on migration flows. The interpretation of the parameters, however, is not straightforward, because the log-linear model is a nonlinear model. In the multiplicative specification of the model, adopted in this article, the parameters are odds and odds ratios. For reasons of interpretation, the multiplicative specification is preferred over the more popular additive formulation, in which the parameters are logits and differences of logits.

To illustrate the combination of various data sources, we considered migration data from the decennial census and the CPS. The decennial census provides detailed information on migration that may be updated using the CPS or, in the future, by turning to its successor, the ACS. The census data are viewed as historical data, and the CPS data are considered to be contemporary data, which at times may consist of judgmental data. The method assures that the model derives its parameters (odds and odds ratios, main effects and interaction effects) from the contemporary data, and when particular odds ratios or interaction effects cannot be derived from the contemporary data (due to lack of detail), they are borrowed from the historical data. That procedure gives priority to contemporary information over historical data.

Preliminary predictions may often be improved upon by the addition of new information. By way of example, we have shown that the prediction of migration from the Northeast to the West, based on a model that combines census data and CPS data, can be improved significantly if judgmental data are added. The judgmental data pertain to qualitative information regarding changes in the interregional migration in the early 1980s that was not captured by the 1980 census or by the data used from the 1985 CPS.

The ACS will generate new opportunities and new challenges, particularly with respect to migration data. It is likely that, to obtain a plausible and consistent picture of migration dynamics, more effective use will need to be made of a variety of sources of migration data. Some sources, such as regional or even

local migration surveys, may need to be introduced to determine changes in the levels and patterns of migration. The method proposed in this article seeks to contribute to such effective uses of different types of information on migration. ■

Literature Cited

Bartlett, M. S. 1935. Contingency table interactions. *Journal of the Royal Statistical Society Supplement* 2:248–52.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Brass, W. 1974. Perspectives in population prediction: Illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society A* 137 (4):532–70.

Coale, A. J., and J. Trussell. 1974. Model fertility schedules: Variations in the age structure of childbearing in human populations. *Population Index* 40 (2):185–258.

Deming, W. E., and F. E. Stephan. 1940. On a least-squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11:427–44.

Fienberg, S. 1970. An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics* 41 (3):907–17. (Corrig. 1971. 42 (5):1778).

Flowerdew, R., and M. Aitkin. 1982. A method for fitting the gravity model based on the Poisson distribution. *Journal of Regional Science* 22:191–202.

Flowerdew, R., and A. Lovett. 1988. Fitting constrained Poisson regression models to interurban migration flows. *Geographical Analysis* 20 (4):297–307.

Fotheringham, A. S., and M. E. O'Kelly. 1989. *Spatial interaction models: Formulations and applications*. Dordrecht: Kluwer Academic Publishers.

Gokhale, D., and S. Kullback. 1978. *The information in contingency tables*. New York: Marcel Dekkar.

Haynes, K. E., and F. Y. Phillips. 1982. Constrained minimum discrimination information: A unifying tool for modeling spatial and individual choice behavior. *Environment and Planning A* 14 (10):1341–54.

Heligman, L., and J. H. Pollard. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries*. part 1 107 (434):49–80.

Isard, W. 1960. *Methods of regional analysis*. Cambridge, MA: MIT Press.

Johnston, R. J., and C. J. Pattie. 1993. Entropy-maximizing and the iterative proportional fitting procedure. *The Professional Geographer* 45 (3):317–22.

Knudsen, D. C. 1992. Generalizing Poisson regression: Including a priori information using the method of offsets. *The Professional Geographer* 44 (2):202–8.

Lin, G. 1999a. Assessing changes in interstate migration patterns of the United States elderly population, 1965–1990. *International Journal of Population Geography* 5 (6):411–24.

———. 1999b. Assessing structural change in U.S. migration patterns: A log-rate modeling approach. *Mathematical Population Studies* 7 (3):217–37.

Plane, D. A. 1981. Estimation of place-to-place migration flows from net migration totals: A minimum information approach. *International Regional Science Review* 6 (1):33–51.

Preston, S. H., P. Heuveline, and M. Guillot. 2001. *Demography: Measuring and modeling population processes*. Oxford: Blackwell.

Rogers, A., ed. 1999. Special issue on the indirect estimation of migration. *Mathematical Population Studies* 7 (3):181–310.

Rogers, A., and L. Castro. 1981. *Model migration schedules*. Research report. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Sen, A., and T. Smith. 1995. *Gravity models of spatial interaction behavior*. Berlin: Springer-Verlag.

Snickars, F., and J. W. Weibull. 1977. A minimum information principle: Theory and practice. *Regional Science and Urban Economics* 7:137–68.

Tyree, A. 1973. Mobility ratios and association in mobility tables. *Population Studies* 27 (3):577–88.

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

U.S. Census Bureau. 1984. *Census of population and housing: 1980*. Public Use Microdata Samples. Washington, DC: U.S. Government Printing Office.

———. 1987. *Geographical mobility: 1985*. Current Population Reports, series P–20, no. 420. Washington, DC: U.S. Government Printing Office.

———. 1993. *Census of population and housing: 1993*. Public Use Microdata Samples. Washington, DC: U.S. Government Printing Office.

Willekens, F. 1983. Log-linear modelling of spatial interaction. *Papers of the Regional Science Association* 52:187–205.

———. 1994. Monitoring international migration flows in Europe: Towards a statistical data base combining data from different sources. *European Journal of Population* 10 (1):1–42.

Willekens, F., A. Por, and R. Raquillet. 1979. *Entropy, multiproportional, and quadratic techniques for inferring detailed migration patterns from aggregate data: Mathematical theories, algorithms, applications, and computer programs*. Working Paper WP–79–88. Laxenburg, Austria: IIASA. Shorter version published 1981 in *IIASA Reports* 4:83–124.

Wilson, A. G. 1970. *Entropy in urban and regional modeling*. London: Pion.

Wong, D. W. S. 1992. The reliability of using iterative proportional fitting. *The Professional Geographer* 44 (3):340–48.

ANDREI ROGERS is a professor of geography and the director of the Population Program at the University of Colorado, Boulder, CO 80309. E-mail: andrei.rogers@colorado.edu. His current teaching and research interests revolve around migration, urbanization, and mathematical demography.

FRANS WILLEKENS is a professor of demography and the head of the Population Research Centre at the University of Groningen, The Netherlands, NL-9700 AV. E-mail: f.j.willekens@frw.rug.nl. His teaching and research interests include spatial demography, statistical modeling, life-history analysis, and forecasting.

JAMES RAYMER is a doctoral student in the Department of Geography and the Population Program at the University of Colorado, Boulder, CO 80309. E-mail: raymer@mail.colorado.edu. His current teaching and research interests include migration, mathematical demography, and statistical modeling.