THE AMERICAN COMMUNITY SURVEY'S INTERSTATE MIGRATION DATA: STRATEGIES FOR SMOOTHING IRREGULAR AGE PATTERNS[1]

James Raymer[2] and Andrei Rogers[3]

1 March 2007 (Draft)

ABSTRACT

Because migrations are relatively rare events, age- and origin-destination-specific flows obtained from population samples often contain irregularities. Bias in analyses of migration flows can arise if these irregularities are not corrected for. In this paper, we present some typical examples of age-specific migration flows with irregular patterns, using the recently released American Community Survey (ACS) data. Strategies for smoothing these patterns based on the multiexponential model migration schedule and the categorical log-linear model are presented and compared. Before smoothing the irregularities found in the ACS data, our ideas are first tested on age-specific interstate migration in the U.S. West Region during 1995-2000 using the 5% Public Use Microdata Sample (PUMS). The corresponding full sample Census data are used to assess the smoothed patterns. Our results demonstrate that more accurate migration data can be provided by smoothing the irregularities caused by relatively small samples.

---

[1] Paper prepared for the 2007 *Annual Meeting of the Population Association of America*, New York, New York.

[2] Division of Social Statistics, School of Social Sciences, University of Southampton

[3] Population Program, Institute of Behavioral Science, University of Colorado

1. INTRODUCTION

Because migrations are relatively rare events, age- and origin-destination-specific flows obtained from population samples often contain irregularities. Bias in analyses of migration flows can arise if these irregularities are not corrected. In this paper, we present some typical examples of age-specific migration flows with irregular patterns, using the recently released American Community Survey data, as well as data from the 2000 Census. Strategies for smoothing these patterns based on the multiexponential model migration schedule (Rogers and Castro 1981) and the categorical log-linear model (Raymer et al. 2006; Raymer and Rogers forthcoming; Rogers et al. 2001; 2002a; 2003) are presented and compared: The model schedule approach is a "bottom-up" approach that smoothes each age-specific flow in a migration flow table. The log-linear model is a "top-down" approach in which higher-order marginal totals of an age-by-origin-by-destination table of migration flows are assumed to be more reliable (and regular) than lower-order marginal totals or cell values. Specific two-way and three-way interaction effects can be dropped to smooth the patterns. Finally, we show how model migration schedules can be incorporated into log-linear models for improvements in both fit and parsimony. To test our ideas, we estimate age-specific interstate migration flows in the U.S. West Region during 1995-2000, using the 5% Public Use Microdata Sample (PUMS) files. These estimates are compared with the corresponding full sample Census data, which are considered to be reliable. We then go on to smooth the irregularities found in the more recent age-specific interstate ACS migration data.

2. MIGRATION DATA FROM THE ACS, 2000-2004

The age-specific patterns of migration from the ACS are examined in this paper. The goal is to identify the more stable structures in these data. The motivation comes from finding many interstate migration flows with irregular age-specific shapes, such as those set out in Figure 1 for migration from California and Washington to other states in the Pacific region (i.e., Alaska, Hawaii, and Oregon). For simplicity, we only examine the migration patterns between these five states and four other regions: Mountain, Northeast, Midwest and South (i.e., a 9-region system with, 72 interregional flows). Both California and Washington are relatively large states. The age-specific patterns get even worse when examining flows from smaller states. There exist many "empty" age-specific cells. Clearly, the ACS migration data breaks down at the interstate level if disaggregated into age groups. Franklin and Plane (2006) were more concerned about the ACS not providing migration data at the county-to-county level. Given the current sampling frame, most of the inter-county data would be unreliable.

---------- Figure 1 about here ----------

A comparison of several age profiles of migration from California to the other four states in the Pacific region is presented in Figure 2. The age profiles come from the 2000 ACS, the 2004 ACS, the aggregated 2000-2004 ACS, and the 2000 Census (full sample). There is a difference between the ACS and Census data in that the former represents one-year migration flows and the latter five-year flows. However, both should have relatively smooth age profiles, yet they do not.

---------- Figure 2 about here ----------

The aggregated ACS data were obtained by summing the counts for the five available ACS years and calculating the age-specific proportions. When aggregated, the flows tend to conform to shapes that one would expect for migration. In fact, they start to resemble the census data. In particular the aggregated CA-OR and CA-WA flows have shapes similar to the corresponding 2000 Census flows. The 2000 and 2004 ACS data, however, are very different and could lead to misleading results if directly adopted for analyses. For example, the 2004 CA-HI flow exhibits a large retirement peak, whereas the 2000 ACS, the aggregated ACS, and the 2000 Census counts do not. Many other large differences appear as well. One usually expects age patterns of migration to evolve gradually over time. For example, Raymer et al. (2006) examined annual interregional migration flows in Italy from 1970 to 2000 obtained from population registers and found strong stability in the patterns over time. Such patterns do not appear in the ACS data.

So, what aspects of the ACS migration data are reliable? Using multiplicative components, we examine various age and spatial structures of migration flows between the five Pacific states and the Mountain, Northeast, Midwest, and South regions from 2000 to 2004. This allows us to identify the more reliable structures of the migration flow data, which can be used to improve migration estimation.

The multiplicative component model for an origin (O) by destination (D) by age (A) table of migration flows is specified as

$$n_{ijx} = (T)(O_i)(D_j)(A_x)(OD_{ij})(OA_{ix})(DA_{jx})(ODA_{ijx}) \qquad i \neq j \qquad (1)$$

where n_{ijx} is an observed flow of migration from origin i to destination j for age group x (i.e., 0-4, 5-9, ..., 80+ years, measured at the beginning of the one-year or five-year time

interval). Note, age-specific proportions (or age compositions) of migration, used later in this paper, are denoted by p_{ijx}, and is calculated as,

$$p_{ijx} = \frac{n_{ijx}}{\sum_{x} n_{ijx}} = \frac{n_{ijx}}{n_{ij+}}. \quad (2)$$

There are eight multiplicative components in total: an overall level, three main effects, three two-way interaction components and a single three-way interaction component. Note, for analysis and estimation purposes, the three-way interaction component ODA_{ijx} is generally ignored because (1) the other seven components capture nearly all of the patterns and (2) because it has a relatively complex interpretation (Raymer et al. 2006).

The components are calculated with reference to the total level in the migration flow tables. The *T* component represents the total number of all migrants in the system,

$$T = \sum_{ijx} n_{ijx} = n_{+++}. \quad (3)$$

The main effect components, O_i, D_j and A_x, represent proportions all migration from each origin, to each destination, and in each age group, respectively, i.e.,

$$O_i = \frac{\sum_{jx} n_{ijx}}{\sum_{ijx} n_{ijx}} = \frac{n_{i++}}{n_{+++}} \quad (4)$$

$$D_j = \frac{\sum_{ix} n_{ijx}}{\sum_{ijx} n_{ijx}} = \frac{n_{+j+}}{n_{+++}} \quad (5)$$

$$A_x = \frac{\sum_{ij} n_{ijx}}{\sum_{ijx} n_{ijx}} = \frac{n_{++x}}{n_{+++}} \quad (6)$$

The two-way interaction components represent the ratios observed migration to expected migration (for the case of no interaction) and are calculated as

$$OD_{ij} = \frac{n_{ij+}}{(T)(O_i)(D_j)} \tag{7}$$

$$OA_{ix} = \frac{n_{i+x}}{(T)(O_i)(A_x)} \tag{8}$$

$$DA_{jx} = \frac{n_{+jx}}{(T)(D_j)(A_x)} \tag{9}$$

These interaction components represent ratios of observed flows or marginal totals to expected ones (i.e., based on the assumption of independence between the variables). The OD_{ij} component captures the association or "connectedness" between origins and destinations. The OA_{ix} and DA_{jx} components represent the deviations from the overall age profile of migration, n_{i+x}. Finally, although not analyzed or estimated in this chapter, the ODA_{ijx} component is calculated as:

$$ODA_{ijx} = \frac{n_{ijx}}{(T)(O_i)(D_j)(A_x)(OD_{ij})(OA_{ix})(DA_{jx})}. \tag{10}$$

The *T*s are set out in Figure 3 for the ACS data from 2000 to 2004. The levels in 2000 and 2004 were roughly the same with 2001, 2002, and 2003 having slightly lower values. From this, we can assume that the overall levels are being captured adequately, as we expect stability over time. As illustrated in Figures 3, 4 and 5, the O_i, D_j and A_x components, respectively, all exhibited stability in their patterns over time. Also the expected patterns appeared, for example, with California sending and receiving the largest share of migrants in Pacific region and Alaska and Hawaii the least. In summary,

from this simple analysis, it appears that the origin, destination, and age main effect components are reasonable and reliable.

---------- Figures 3, 4 and 5 about here ----------

Finally, the OD_{ij} components are illustrated over time in Figure 6 for migration flows from California and Washington. Here, there appears to be some stability over time in the patterns, though not as strong as those found in the main effect components. However, they do correspond to what one would expect, say, for migration from California and Washington. That is, we expect to find strong associations in the migration patterns between neighboring states or regions and weak associations between non-neighboring states or regions. For this paper, we assume that these associations are reliable. This implies the aggregate interstate migration levels of the ACS can be trusted, leaving us to focus on the age profiles of migration which are clearly not reliable. Future research should explore the reliability of the aggregate origin-destination-specific flows.

---------- Figure 6 about here ----------

3. SMOOTHING METHODS

The strategies, strengths, and weaknesses of fitting model migration schedules and log-linear models are set out in this section.

3.1 Model Migration Schedules

Model migration schedules (Rogers & Castro 1981) are used in this paper to smooth age profiles of migration, p_{ijx}, under the assumption that the aggregate origin-destination-specific flows, n_{ij+}, are reasonable and more reliable than the corresponding disaggregate

flows, n_{ijx}. With this assumption, we can obtain smoothed age-specific migration flows by positing that $\hat{n}_{ijx} = \hat{p}_{ijx} n_{ij+}$, where $\hat{n}_{ijx}$ and $\hat{p}_{ijx}$ denote predicted age-specific flows and proportions, respectively, of interregional migration.

We use a seven-parameter model migration schedule to smooth the age profiles of migration. This model consists of three components: a constant minimum level of migration, a negative exponential curve that represents child migrant flows, and a double-exponential curve that represents the young adult migrants around the age of the "labor force peak".

$$\hat{p}_{ijx} = a_0 + a_1 \exp(-\alpha_1 x) + a_2 \exp\{-\alpha_2 (x - \mu_2) - \exp[\lambda_2 (x - \mu_2)]\}, \qquad (11)$$

where the *a*'s denote level parameters and the α's, μ's, and λ represent shape parameters. In examining data from the 2000 Census, retirement peaks were not found in the age profiles of age-specific migration between the regions covered in this paper. Therefore we simply apply the above model schedule; Schedules with retirement peaks can be modeled by adding another (four-parameter) double-exponential curve to the seven-parameter model migration schedule equation (Rogers & Castro 1981).

3.2 Log-Linear Models

Log-linear models have been used recently to analyze age and spatial structures of migration (Raymer et al. 2006; Raymer and Rogers forthcoming; Rogers et al. 2002). These models use maximum likelihood methods for parameter estimation and assume that the counts are Poisson distributed. The saturated model is specified as

$$\log n_{ijx} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA} + \lambda_{ijx}^{ODA}. \qquad (12)$$

Migration flow tables can be smoothed by dropping various two-way or three-way interaction terms. For example, the unsaturated model with structural zeros for non-migrants (i.e., n_{ii}):

$$\log \hat{n}_{ijx} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \delta_i I(i \neq j). \quad (13)$$

provides estimates of migration flows that are consistent the observed aggregate levels (i.e., n_{ij+}) but have a single age profile of migration (i.e., p_{++x}) applied to all flows. Structural zeros are included through the $\delta_i I(i \neq j)$ term (Agresti 2002), in this case for cells representing non-migrants or intraregional migrants. Note, $I(\cdot)$ is an indicator function, $I(i \neq j) = \begin{cases} 1, i \neq j; \\ 0, i = j \end{cases}$.

3.3 Incorporating Model Migration Schedules in Log-Linear Models

The motivation for this third approach comes from recent work on estimating age-specific migration flows in the context of internal migration in the U.S. (Raymer and Rogers forthcoming) and international migration in Europe (Raymer forthcoming). Various structures can be included with unsaturated models via offsets. For example, the model,

$$\log \hat{n}_{ijx} = \log n_{ijx}^{*} + \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD}. \quad (14)$$

provides estimates of migration flows that are consistent the observed aggregate levels of flows (i.e., n_{ij+}) but borrow age profiles of migration from the offset or auxiliary data set, n_{ijx}^{*}. Structural zeros also can be included in the offset to remove cells representing non-migrants or intraregional migrants from the estimation. With the offset, model migration

schedules of the reported data or of the aggregate flows of in-migration or out-migration can be incorporated to improve the prediction of age-specific interregional migration flows. The advantage is that the predicted flows will be forced to fit the margins of the origin-destination-age migration table that are believed to be reliable.

In the next section, we test the models in Section 3.1 and Section 3.2 on the U.S. 2000 Census data. For these data we have both the full sample and the PUMS 5% sample, with the latter exhibiting more irregularities than the former (as expected). We compare the smoothed estimates of the PUMS data against the full sample data using the R^2 goodness-of-fit test.

4. SMOOTHING THE 5% CENSUS 2000 SAMPLE

In this section, we test the above ideas on the 5% Public Use Microdata Sample (PUMS) of the 2000 Census. These data have some of the problems visible in the ACS, although they are not as prevalent or as significant. Because the PUMS 5% data is relatively good and captures most of the full-sample age-specific interstate migration patterns, we expand our analysis to migration between twenty states or regions, that is, migration between the thirteen states in the West region and the seven divisions outside the West region (i.e., New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, and West South Central). Furthermore, we focus on migration from Colorado because of its relatively low levels of migration and consequently more irregular patterns. For this section, we first compare the age profiles of migration from Colorado to the twelve other states in the West region and to the seven divisions outside the West region obtained from the 2000 Census full sample and the 2000 Census PUMS.

Second, model migration schedules are fitted to the PUMS data. Third, the unsaturated log-linear model with all two-way interactions included,

$$\log \hat{n}_{ijx} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA} + \delta_i I(i \neq j), \qquad (15)$$

is used to smooth the age profiles of the PUMS data. The two sets of predicted age- and origin-destination specific flows from Colorado are compared to each other and with the full sample Census data.

Since there were no elderly peaks apparent in the nineteen age-specific migration flows from Colorado, we applied the seven parameter model migration schedule (Equation 1) to smooth the age profiles (standardized to unit area). *Results are presented in Table 1 and Figure 7. Discussion to be continued...*

---------- Table 1 and Figure 7 about here ---------

5. SMOOTHING THE IRREGULARITIES CONTAINED IN THE 2004 AMERICAN COMMUNITY SURVEY MIGRATION DATA

The ACS data are smoothed in this section, using the geography set out in Section 2. First we fit model migration schedules to the 2004 age profiles of migration (i.e., p_{ijx}). Second, we compare these fits with the estimates obtained from the unsaturated log-linear model with all two-way interactions included. Finally, we combine model migration schedules and log-linear models using the approach set out in Section 3.3. This is necessary because in the 2004 ACS data, even the aggregate age-specific inflows and outflows contain irregularities.

There are four age-specific flows set out in Figure 8 representing a combination of ACS migration flow data. The flow from Hawaii to Alaska is a situation where there

are only seven data points. The flows from Hawaii to California and California to Oregon are cases where the patterns are highly irregular. And, the flow from Washington to Oregon contains an age profile that is fairly regular with the exception of a small peak at the 50-54 age group. We compare three predictions: model migration schedules, unsaturated log-linear model with all two-way interactions included, and a hybrid log-linear model that combines smoothed model schedule fits of aggregate in-migration and out-migration.

Out of the 72 flows, only 8 were deemed very difficult or impossible to fit model schedules to. These were AK-HI, AK-OR, AK-NE, HI-OR, OR-HI, WA-HI, NE-OR, and the MW-OR flows.

The main problem with the unsaturated log-linear model predictions is the carrying forward of irregular patterns contained in the marginal totals.

We use the log-linear-with-offset model specified in Equation 5 to estimate the patterns. OA_{ix} and DA_{jx} were estimated by dividing the model scheduled inflows and outflows (i.e., $\hat{p}_{i+x}$ and $\hat{p}_{+jx}$) by A_x (or p_{++x}). The initial estimated values were obtained by multiplying all the components set out in Equation one, except the three-way interaction component between origin, destination, and age (i.e., ODA_{ijx}). The resulting estimates were then used as the offset, n^*_{ijx}.

Some selected results are presented in Figure 8. Clearly the unsaturated model is inappropriate given that the marginal structures are irregular. Model schedules have the advantage of making the most use out of the reported data but involve a large amount of work and do not work when the data are very irregular. The hybrid log-linear model works well but will miss some of the distinct age patterns that differ from those contained

in the marginal structures (e.g., the Washington to Oregon flow). This model is better equipped to assign an age profile when the data are highly irregular or impossible to fit model schedules to.

---------- Figure 8 about here ----------

6. DISCUSSION

Key things: smoothing improves sample estimates, age-specific ACS migration data are very rough and do not exhibit stability over time, models are needed to improve estimates. The ones we present are relatively simple but effective. *Discussion to be continued...*

REFERENCES

Agresti A. 2002. *Categorical data analysis*. Hoboken: Wiley-Interscience.

Franklin RS and Plane DA. 2006. Pandora's box: The potential and peril of migration data from the American Community Survey. *International Regional Science Review* 29(3):231-246.

Raymer J. forthcoming. Obtaining an overall picture of population movement in the European Union. In *The estimation of international migration in Europe: Issues, models and assessment*, Raymer J and Willekens F, eds. Chichester: Wiley.

Raymer J, Bonaguidi A and Valentini A. 2006. Describing and projecting the age and spatial structures of interregional migration in Italy. *Population, Space and Place* 12:371-388.

Raymer J and Rogers A. forthcoming. Using age and spatial flow structures in the indirect estimation of migration streams. *Demography*.

Rogers A and Castro LJ. 1981. Model migration schedules. RR-81-30, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Rogers A, Willekens FJ, Little JS and Raymer J. 2002a. Describing migration spatial structure. *Papers in Regional Science* 81:29-48.

Rogers A, Willekens FJ and Raymer J. 2001. Modeling interregional migration flows: Continuity and change. *Mathematical Population Studies* 9:231-263.

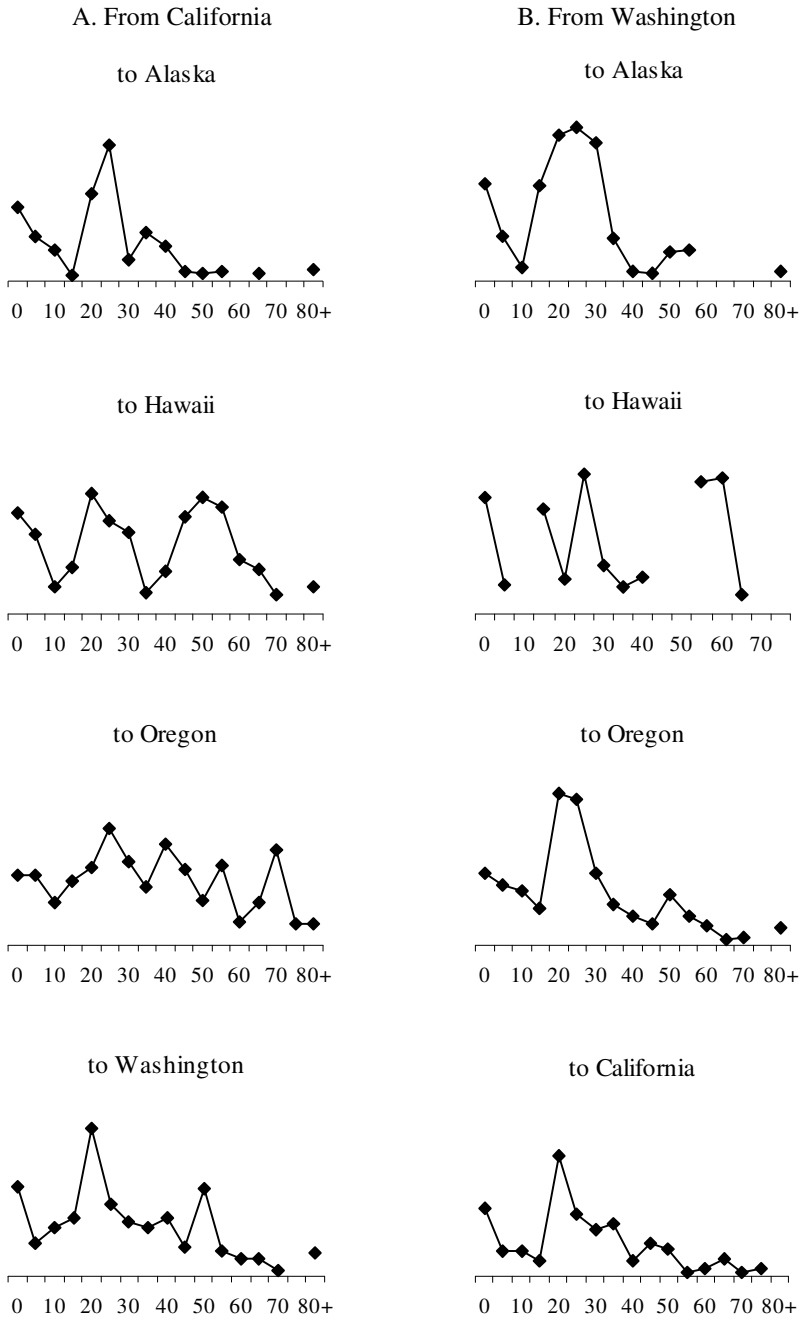
Rogers A, Willekens FJ and Raymer J. 2002b. Capturing the age and spatial structures of migration. *Environment and Planning A* 34:341-359.

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

Table 1. Goodness-of-fit tests (R^2, coefficient of determination) of age compositions of migration from Colorado, Census 2000 data, 1995-2000

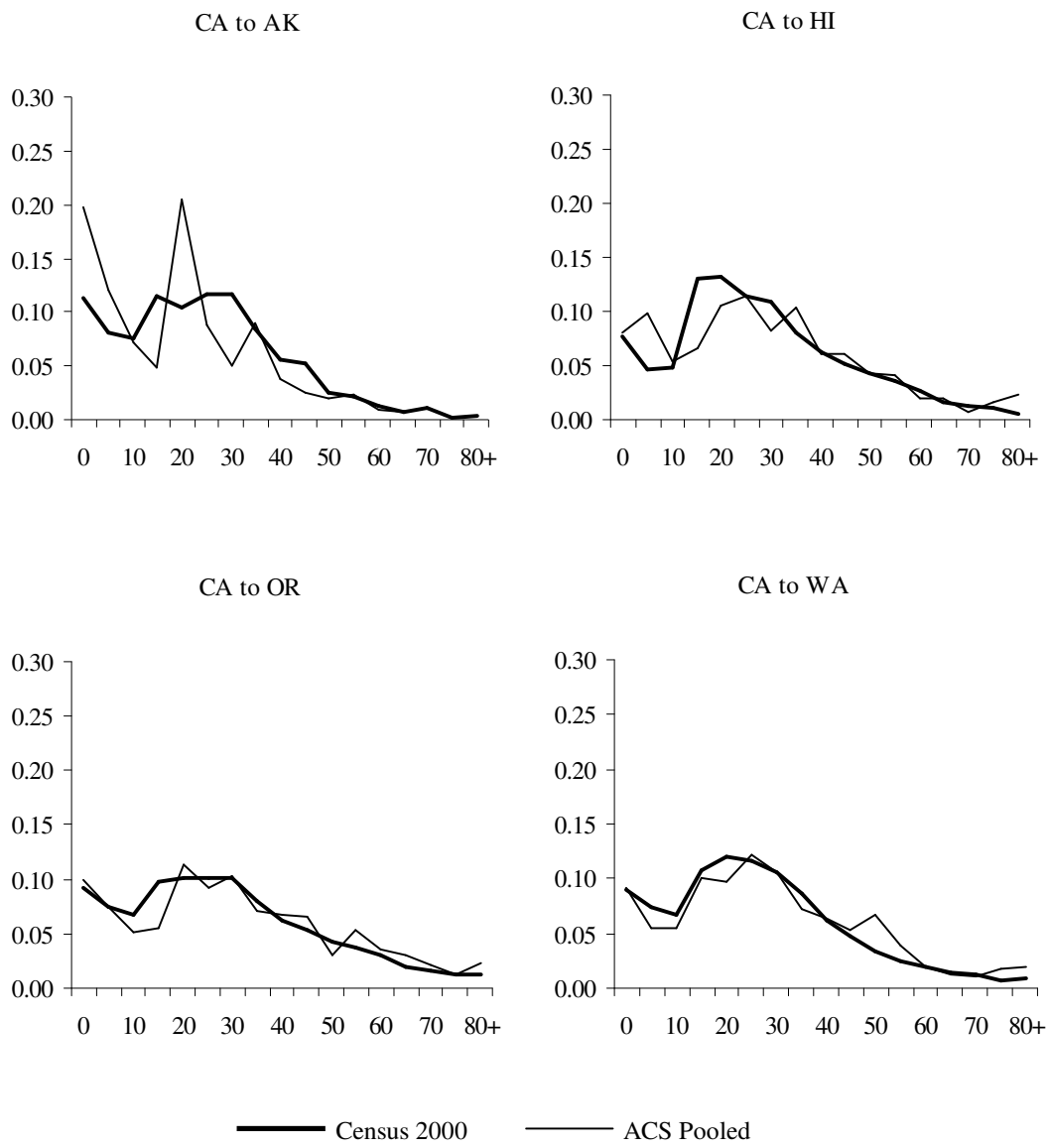
Destination	5% PUMS	Model Migration Schedule	Log-Linear Model
AK	0.942	0.885	**0.975**
AZ	**0.972**	0.882	0.940
CA	0.988	0.983	**0.998**
HI	0.856	0.878	**0.935**
ID	0.858	0.955	**0.958**
MT	0.774	0.809	**0.952**
NV	0.938	**0.970**	0.905
NM	0.940	0.936	**0.944**
OR	0.972	0.968	**0.980**
UT	0.975	0.951	**0.983**
WA	**0.993**	0.980	0.977
WY	0.834	**0.950**	0.916
N.E.	0.980	**0.985**	0.975
M.A.	0.989	**0.990**	0.976
E.N.C.	0.977	0.976	**0.996**
W.N.C.	0.984	**0.992**	0.976
S.A.	0.994	**0.997**	0.970
E.S.C.	0.977	**0.983**	0.974
W.S.C.	**0.993**	0.979	0.989
Average	0.944	0.950	**0.964**
Min	0.774	0.809	0.905
Max	0.994	0.997	0.998
STDev	0.065	0.051	0.026

Notes: (1) Predicted values are compared to full sample 2000 Census data. (2) Best fits are set boldface.



Note: y-axis = level (proportion) and x-axis = age

Figure 1. Interstate migration age compositions from California and Washington to the other states in the Pacific Division, ACS data, 2004



Notes: (1) AK = Alaska, CA = California, HI = Hawaii, OR = Oregon, and WA = Washington; (2) y-axis = proportions; (3) x-axis = age.

Figure 2. A comparison of interstate migration age compositions from California to the other states in the Pacific region: ACS 2000-2004 pooled (one-year interval) and Census 2000 full sample (five-year interval)

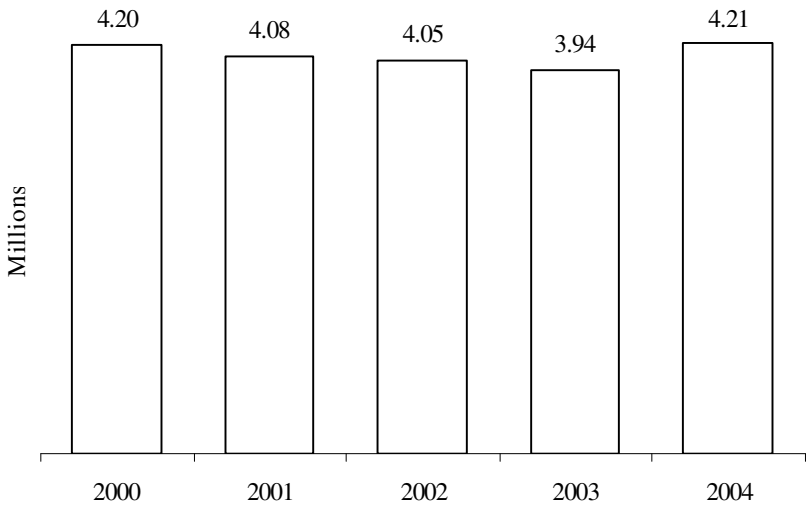


Figure 3. Overall levels of migration between states in the Pacific and the Mountain, Northeast, Midwest, and South regions, ACS data, 2000-2004

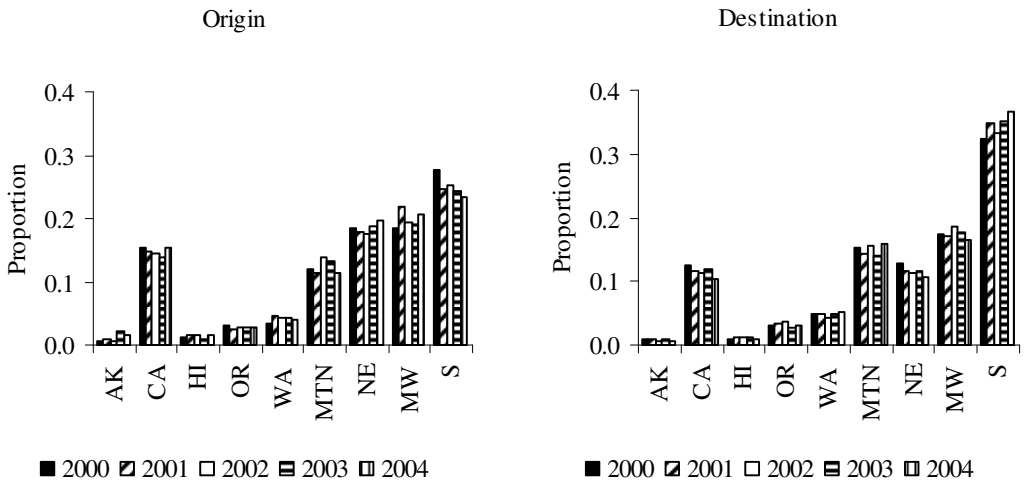


Figure 4. Proportions of all migrants from and to states in the Pacific and the Mountain, Northeast, Midwest, and South regions, ACS data, 2000-2004

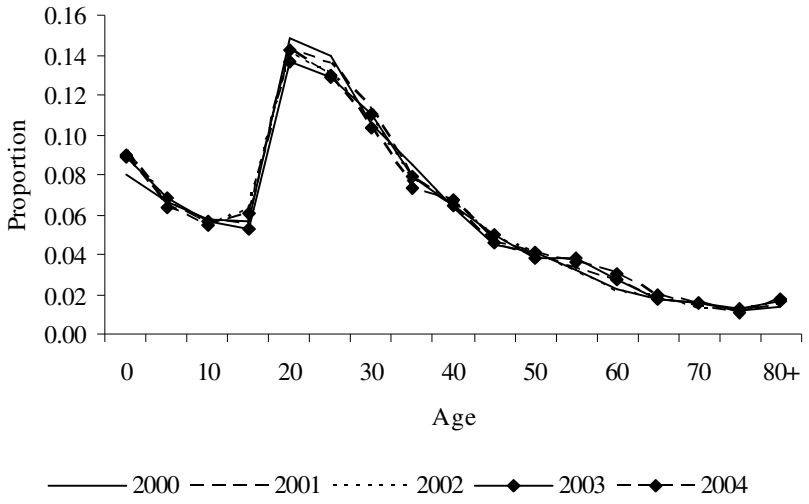
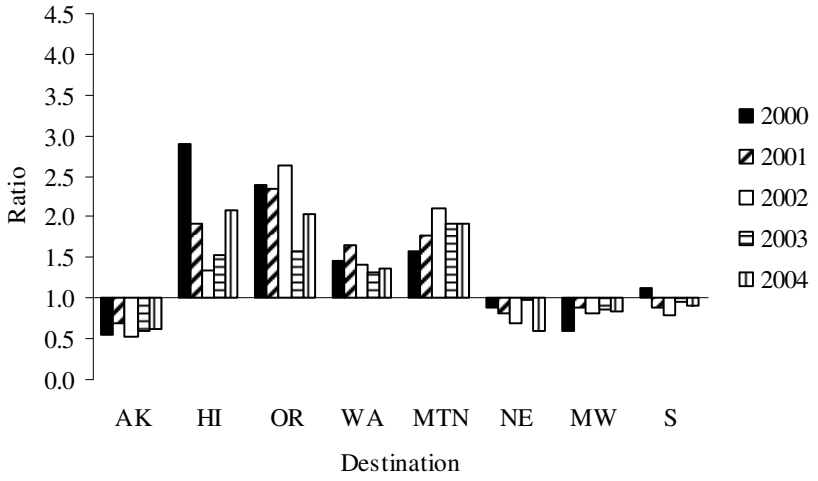


Figure 5. Proportions of migration by age between states in the Pacific and the Mountain, Northeast, Midwest, and South regions, ACS data, 2000-2004

A. From California



B. From Washington

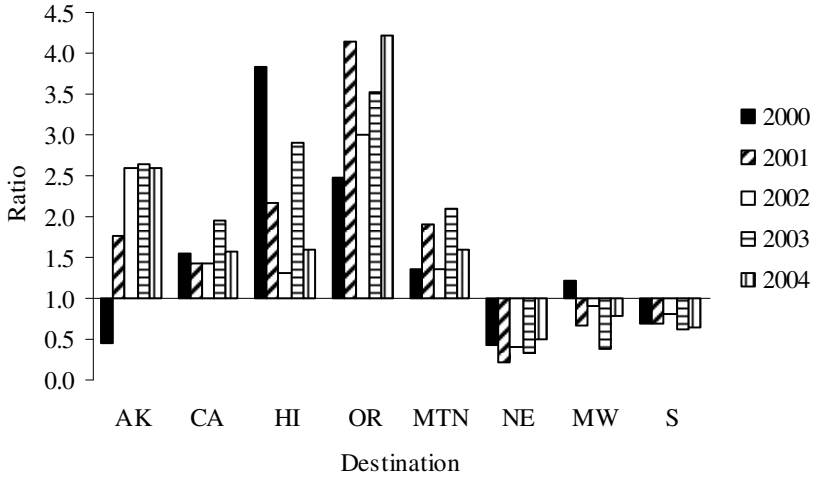
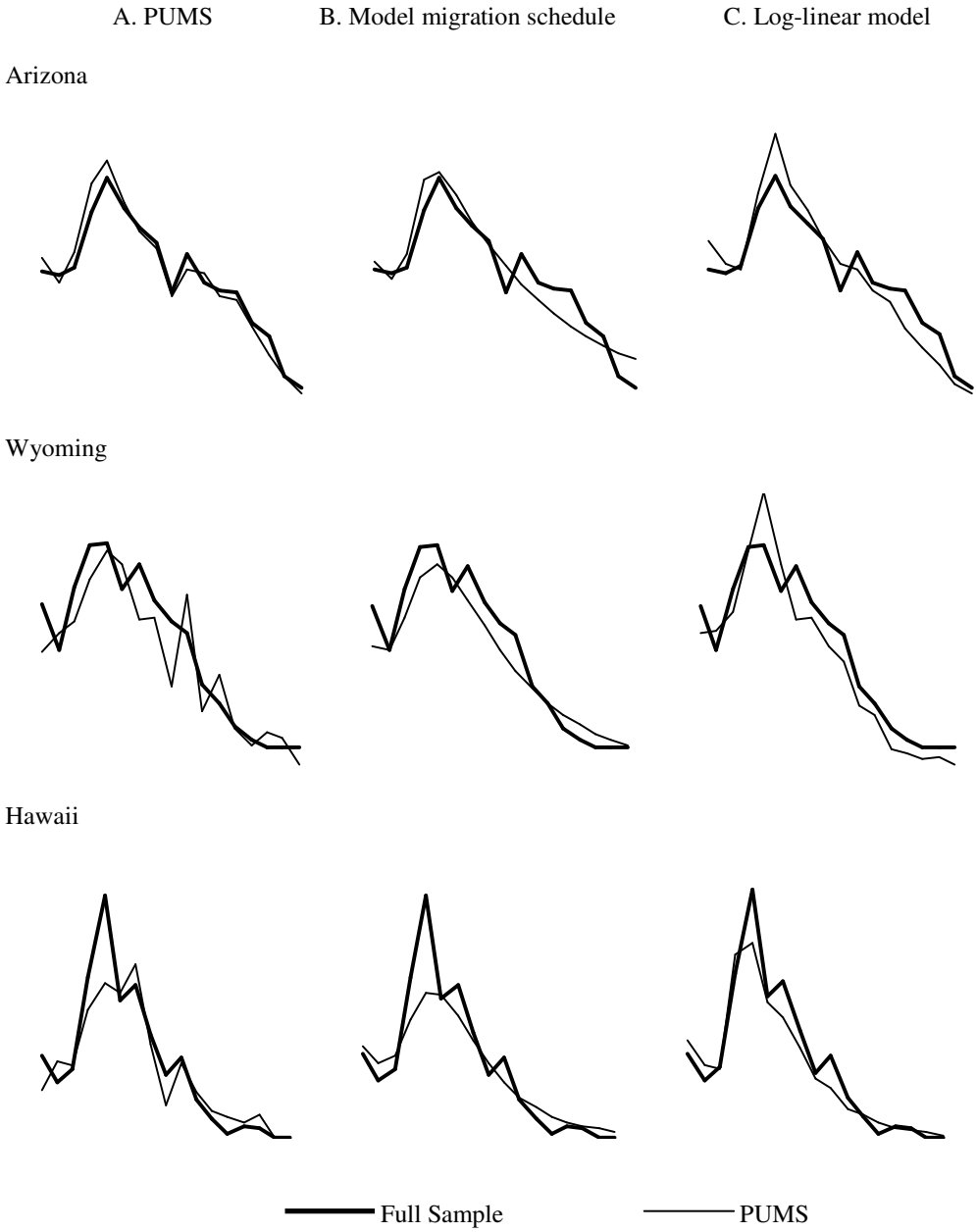


Figure 6. Origin-destination associations of migration from California and Washington to states in the Pacific and the Mountain, Northeast, Midwest, and South regions, ACS data, 2000-2004



Notes: (1) y-axis = level (count) and x-axis = age. (2) See Table 1 for goodness-of-fits.

Figure 7. Observed and predicted age-specific migration flows from Colorado to Arizona, Wyoming and Hawaii, Census 2000 data, 1995-2000 flows: PUMS 5%, model migration schedule fits of PUMS 5% data, and unsaturated log-linear model fits of PUMS 5% data

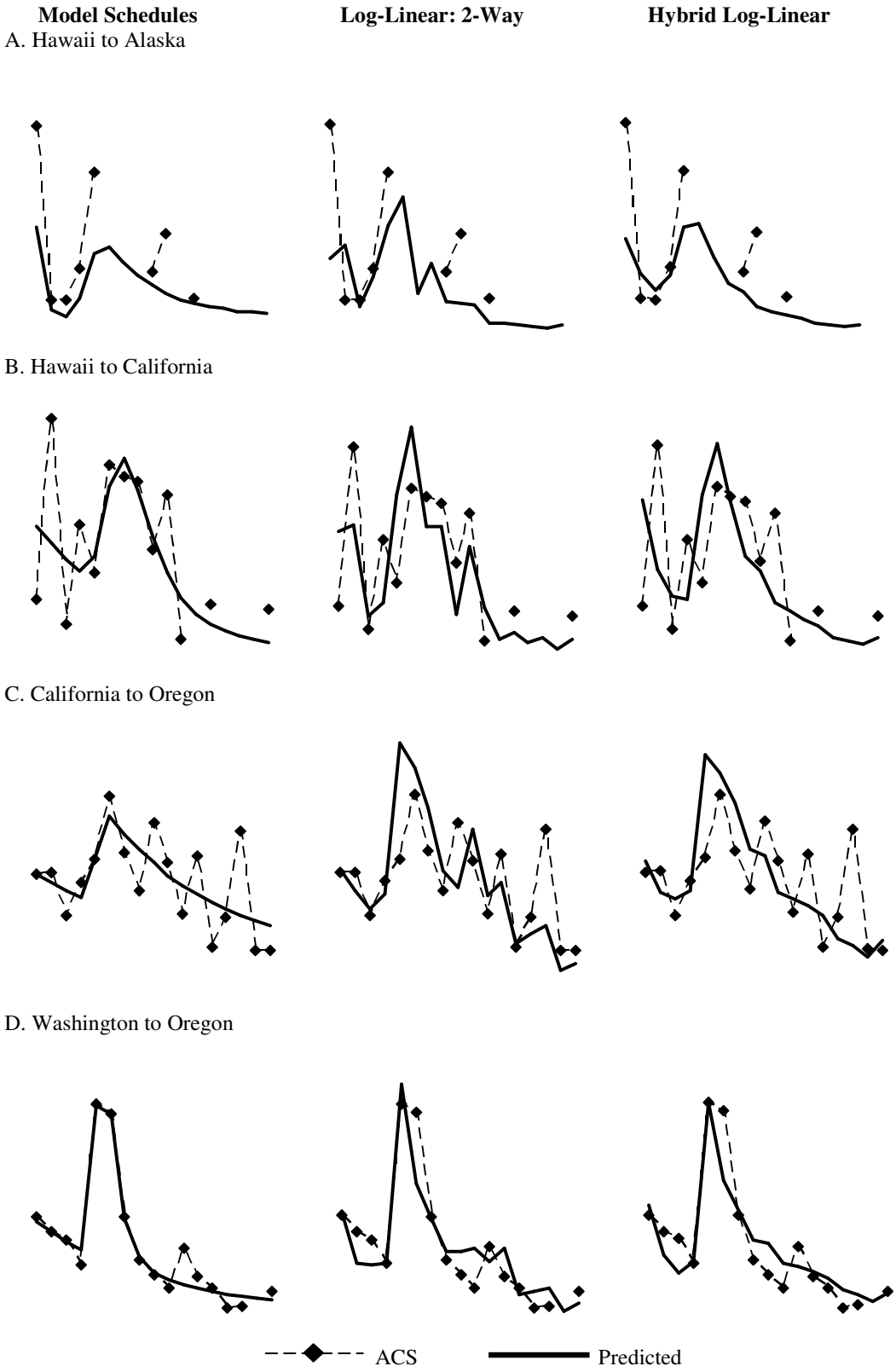


Figure 8. Selected ACS and predicted age-specific migration flows, 2004: Model migration schedules, unsaturated log-linear model, and hybrid log-linear model