Inferring Migration from Birthplace-Specific Population Stocks

Andrei Rogers, James Raymer, Lisa Jordan, and Jun-wei Liu

Population Program
Institute of Behavioral Science
University of Colorado
Boulder, Colorado 80309-0484

Acknowledgments: This research is being supported by a grant from the National Science Foundation (SES-0240808).

Introduction to Method 1: Infant Population Stocks

September 24, 2004

The Importance of Indirect Estimation of Migration Flows

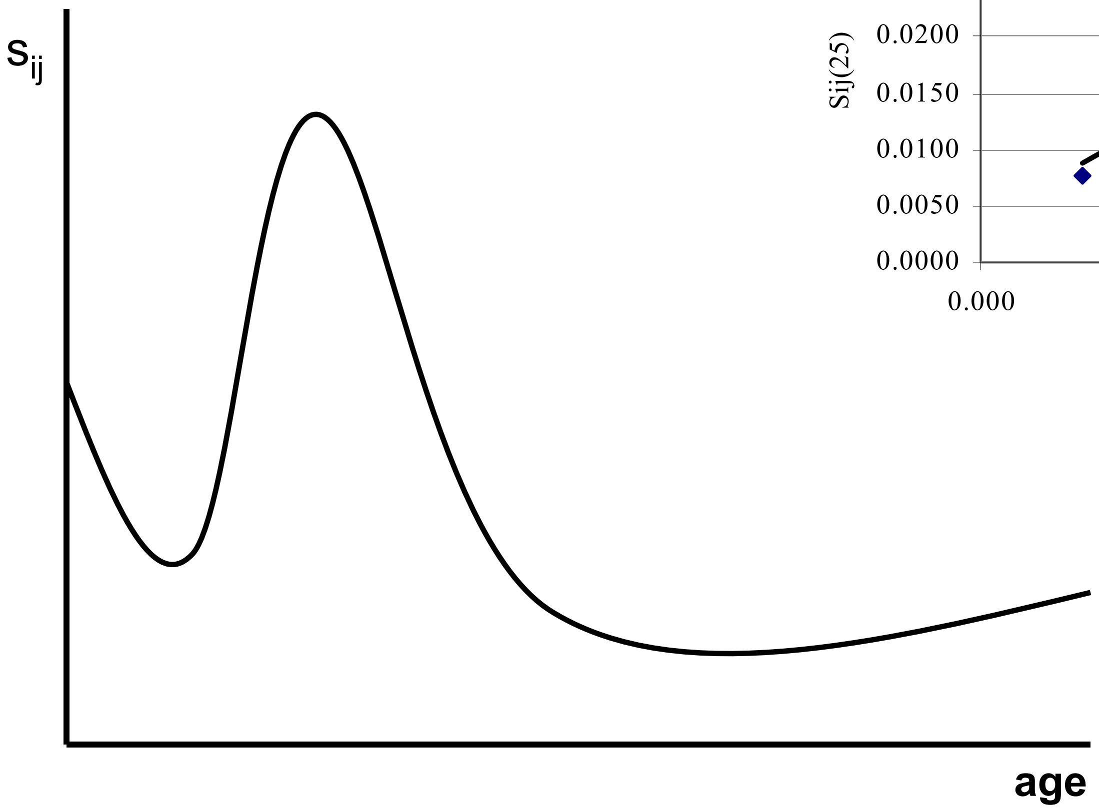
- Indirect estimation is useful for generating information on age-specific, inter-regional migration flows, when no data on migration flows are collected. These data are interesting of themselves, and also valuable for predicting population change, and for use in government and business planning.

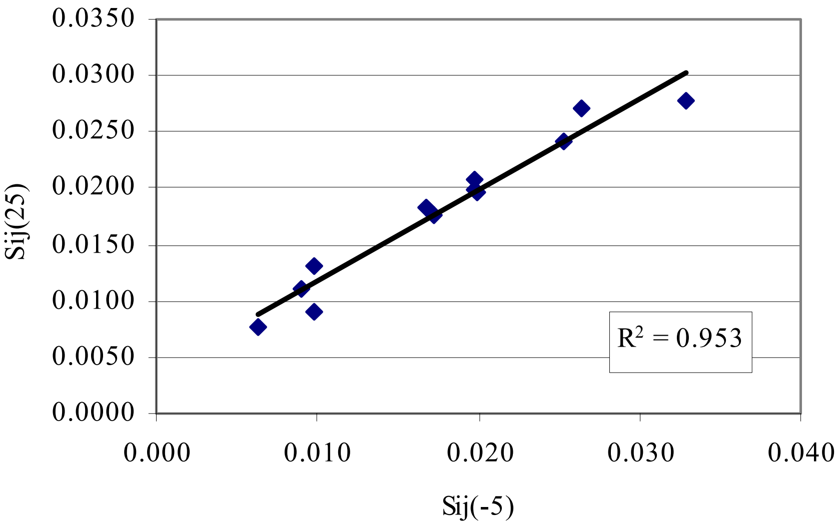
<u>**Overview of Method 1**</u>**:** Inferring Migration Flows from Infant Population Stocks

- Assumptions of the method
- Outline of the method
- Data examined in this case study
- An example of implementing method 1 in EXCEL
- How the results can be assessed
- Summary of the results
- Suggestions for improving the results

ASSUMPTIONS:

- Model Migration Schedule





Inferring Survivorship Proportions from Infant Population Stocks

For this project, the migration values being estimated are survivorship proportions, which are equal to the number of migrants moving from *i* to *j* divided by the total population in *i*.

$$S_{ij}(x) = \frac{Migrants_{ij}(x)}{TotalPop_i(x)}, x = age$$

Another interpretation of survivorships, a measurement taken from mortality studies, is that an individual in *i*, has the probability of surviving ($S_{ij}(x)$) in *j*, *t* years later.

The calculation of infant survivorship proportions only requires data on the population stocks, which makes this an attractive method.

Inferring survivorships takes place by using the age-to-infant ratio

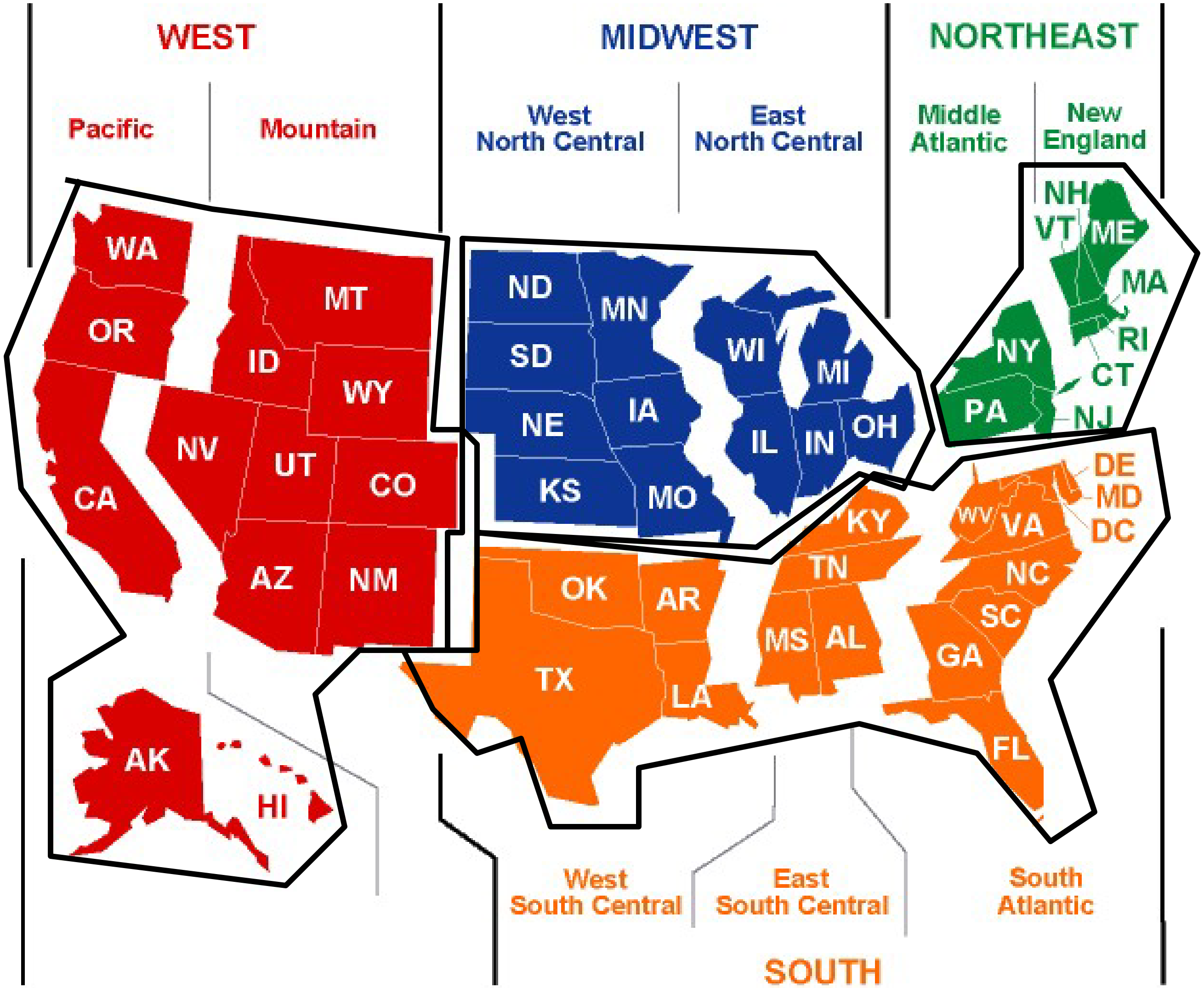
Age-to-Infant Ratio:

$$r_{ij}(x,-5) = \frac{S_{ij}(x)}{S_{ij}(-5)}, \qquad x = 0, 5, 10, \ldots, 80$$

$$\widehat{S}_{ij}^{t}(x) = r_{ij}^{t-10} S_{ij}^{t}(-5), \qquad x = 0, 5, 10, \ldots, 80$$

Data

- U.S. Census 1985-90 data is used to predict 1995-2000, four-region, interregional migration for U.S. born migrants
- The same information will be examined in all three methods for comparison.



An Illustration of Indirect Estimation

Suppose you have data for 1985-90 across a number of age groups. Suppose you also have data to create survivorships for the first age category for 1995-2000.

With this information you can infer the survivorships that you are missing, given that data that you do have.

Year	Origin	Dest	sij(-5)	sij0	sij5	sij10	sij15	sij20	sij25	sij30	sij35	sij40
1985-1990	NE	MW	0.009728	0.006911	0.010432	0.015763	0.014685	0.011378	0.008983	0.008033	0.005912	0.004673
		S	0.025354	0.023295	0.028052	0.042168	0.037612	0.027367	0.024176	0.021581	0.020249	0.018944
		W	0.009852	0.008368	0.009959	0.021545	0.023460	0.016807	0.013123	0.010119	0.007488	0.006501
Year	Origin	Dest	sij(-5)	sij0	sij5	sij10	sij15	sij20	sij25	sij30	sij35	sij40
1990-2000	NE	MW	0.009400									
		S	0.025617									
		W	0.009546									

An Illustration of Indirect Estimation

First, you will need to generate the ATI ratios for the years that you do have data. Recall that this is done simply by dividing: $S_{ij}(x)/S_{ij}(-5)$.

Year	Origin	Dest	sij(-5)	sij0	sij5	sij10	sij15	sij20	sij25	sij30	sij35	sij40
1985-1990	NE	MW	0.009728	0.006911	0.010432	0.015763	0.014685	0.011378	0.008983	0.008033	0.005912	0.004673
		S	0.025354	0.023295	0.028052	0.042168	0.037612	0.027367	0.024176	0.021581	0.020249	0.018944
		W	0.009852	0.008368	0.009959	0.021545	0.023460	0.016807	0.013123	0.010119	0.007488	0.006501
Year	Origin	Dest	rij(-5)	rij0	rij5	rij10	rij15	rij20	rij25	rij30	rij35	rij40
1985-90	NE	MW	1.000000	0.710391	1.072355	1.620319	1.509486	1.169584	0.923378	0.825734	0.607732	0.480365
		S	1.000000	0.918769	1.106403	1.663148	1.483441	1.079372	0.953539	0.851171	0.798638	0.747176
		W	1.000000	0.849388	1.010834	2.186794	2.381163	1.705893	1.331958	1.027126	0.760005	0.659840
Year	Origin	Dest	sij(-5)	sij0	sij5	sij10	sij15	sij20	sij25	sij30	sij35	sij40
1990-2000	NE	MW	0.009400									
		S	0.025617									
		W	0.009546									

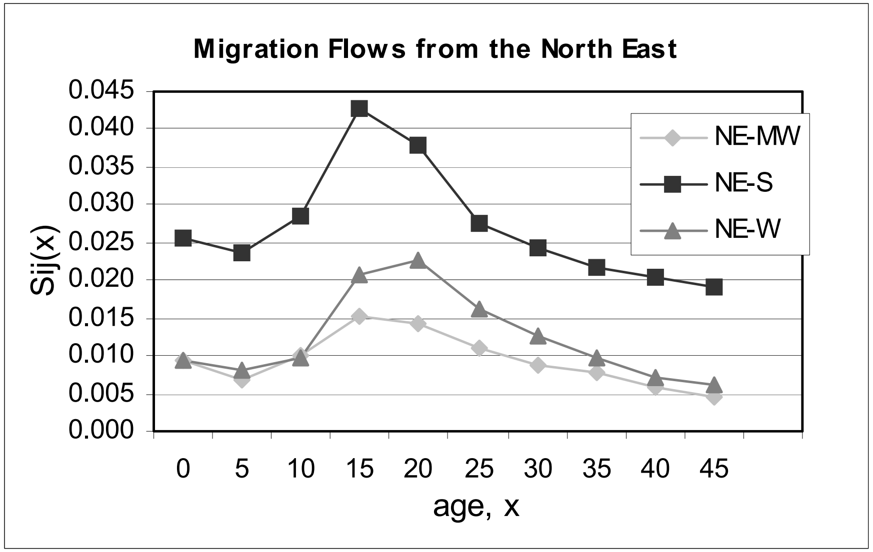
An Illustration of Indirect Estimation

Now, given the ATI ratios, you can predict the survivorships for the age groups that you are missing. This is done by multiplying the ATI from the previous year by the survivorship for the first age group for the year with missing data.

Year	Origin	Dest	rij(-5)	rij0	rij5	rij10	rij15	rij20	rij25	rij30	rij35	rij40
1985-90	NE	MW	1.000000	0.710391	1.072355	1.620319	1.509486	1.169584	0.923378	0.825734	0.607732	0.480365
		S	1.000000	0.918769	1.106403	1.663148	1.483441	1.079372	0.953539	0.851171	0.798638	0.747176
		W	1.000000	0.849388	1.010834	2.186794	2.381163	1.705893	1.331958	1.027126	0.760005	0.659840

Year	Origin	Dest	sij(-5)	sij0	sij5	sij10	sij15	sij20	sij25	sij30	sij35	sij40
1990-2000	NE	MW	0.009400	0.006678	0.010080	0.015231	0.014189	0.010994	0.008680	0.007762	0.005713	0.004515
		S	0.025617	0.023536	0.028342	0.042604	0.038001	0.027650	0.024426	0.021804	0.020458	0.019140
		W	0.009546	0.008108	0.009649	0.020875	0.022730	0.016284	0.012715	0.009805	0.007255	0.006299

You can then plot the data you have created. In this study, we have done the following for 4 regions and destinations. Since we do have data for 1995-2000, we can assess the quality of the estimates.



How to Assess Predicted Values

Mean Algebraic Percent Error (MALPE):

$$\sum_{ij} \left(\frac{\hat{S}_{ij}^{2000}(x) - S_{ij}^{2000}(x)}{S_{ij}^{2000}(x)} \right)$$

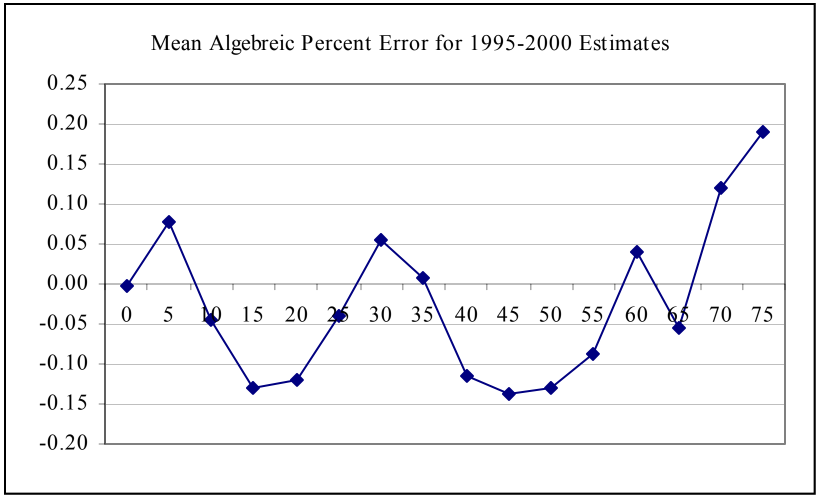
Mean Absolute Percent Error (MAPE):

$$\sum_{ij} \left(\frac{\left| \hat{S}_{ij}^{2000}(x) - S_{ij}^{2000}(x) \right|}{S_{y}^{2000}(x)} \right)$$

R-squared Values for Each Migration Flow

Assessment of Predicted Values, 1995-2000

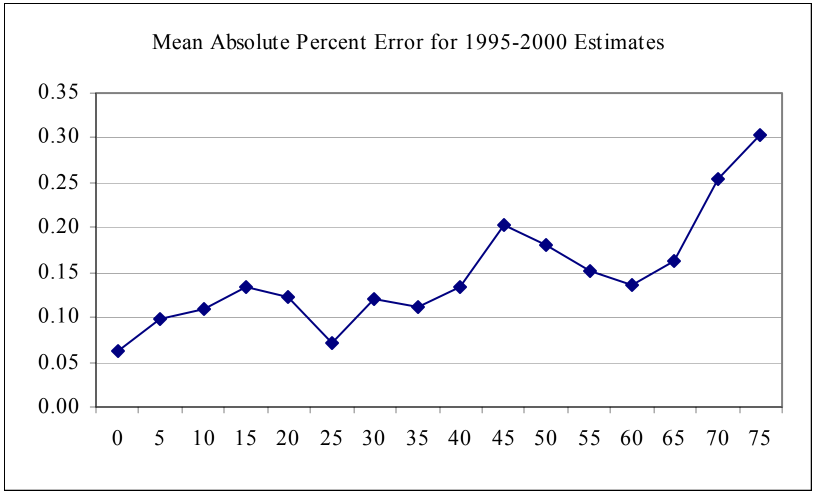
Mean Algebraic Percent Error = -2.3%



The Mean Algebraic Percent Error shows where the predicted values over and under-estimate the actual values.

The age-to-infant ratio best predicts ages 0-5 and 25-29. It tends to most overestimate ages 30-34, 60-64, and 70+, and underestimate 15-24 and 40-59.

Mean Absolute Percent Error = 15%

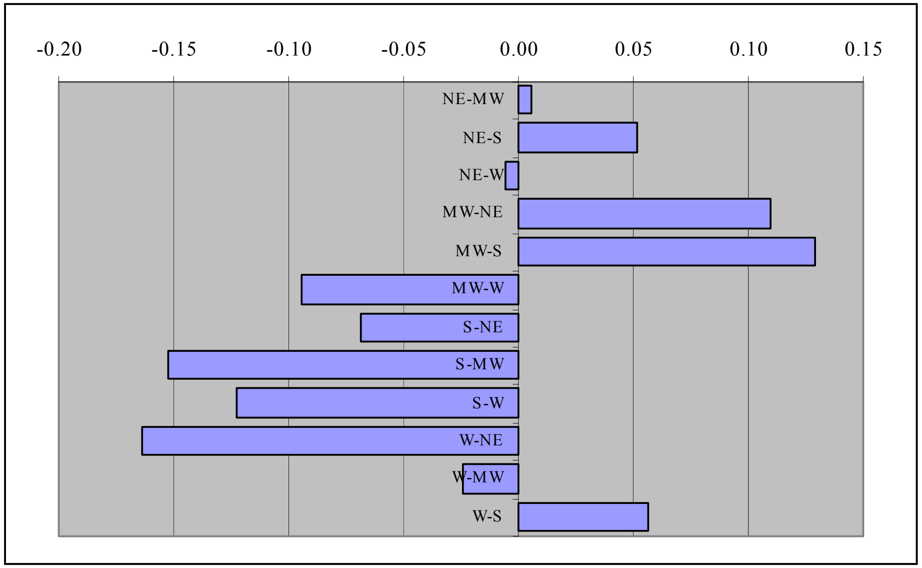


The Mean Absolute Percent Error shows how the overall level of error varies over the age categories.

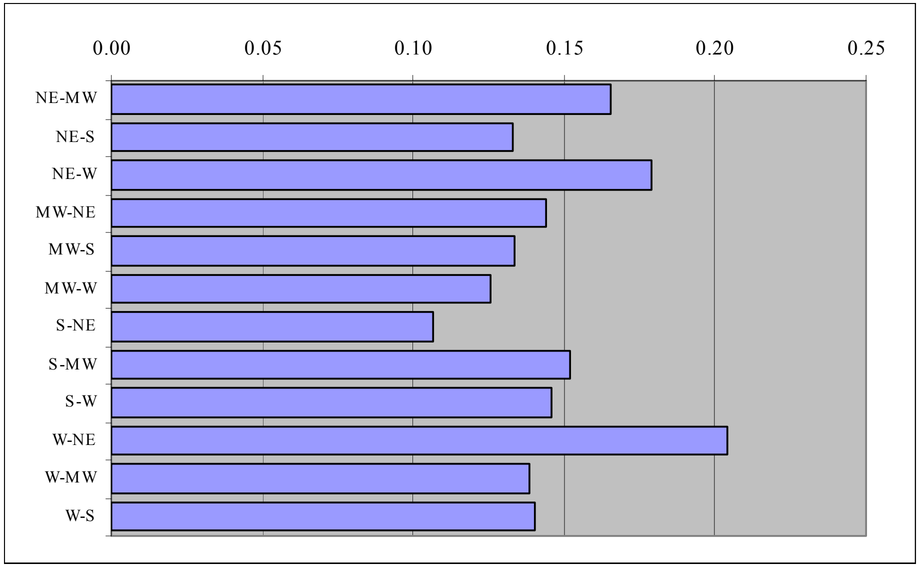
The error is the greatest for ages 75-79, where MAPE = 30%, and lowest for ages 0-5, where the MAPE = 6.2%.

Assessment of Predicted Values, 1995-2000

Mean Algebraic Percent Error by Region



Mean Absolute Percent Error by Region



The Mean Algebraic Percent Error shows where the predicted values over and under-estimate the actual values.

The age-to-infant ratio best predicts flows from W-MW. It tends to most overestimate flows from NE-S and underestimate flows from S-MW.

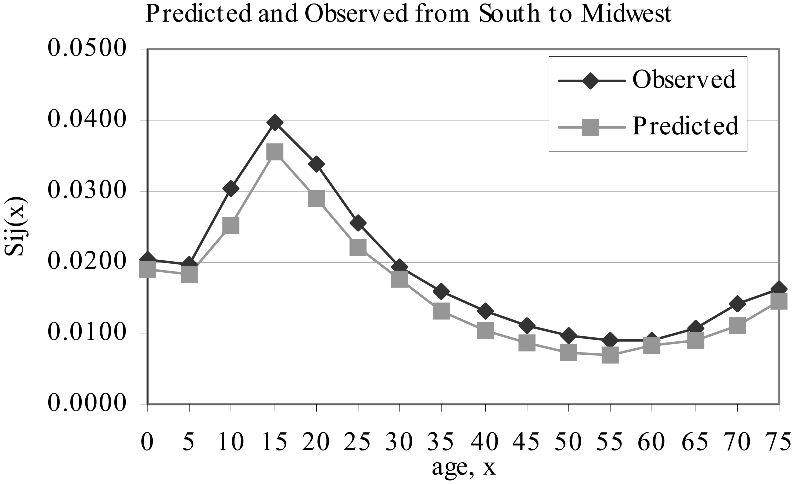
The Mean Absolute Percent Error shows how the overall level of error varies over the different flows.

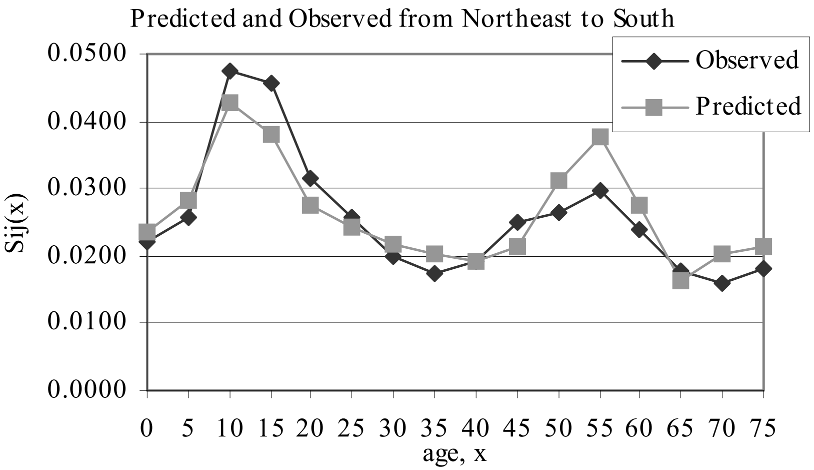
The error is the greatest for the Northeast to South, and lowest for the South to Northeast flows.

Assessment of Predicted Values, 1995-2000

R-squared Values by Region

Year	Origin	Dest	R-sqr
1995-2000	NE	MW	0.872
		S	0.809
		W	0.933
	MW	NE	0.977
		S	0.928
		W	0.937
	S	NE	0.952
		MW	0.989
		W	0.973
	W	NE	0.972
		MW	0.951
		S	0.962

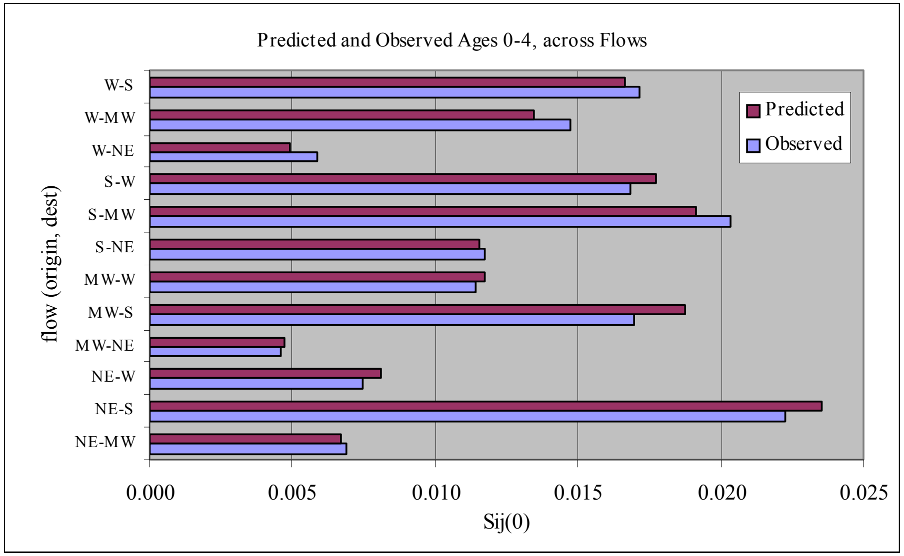


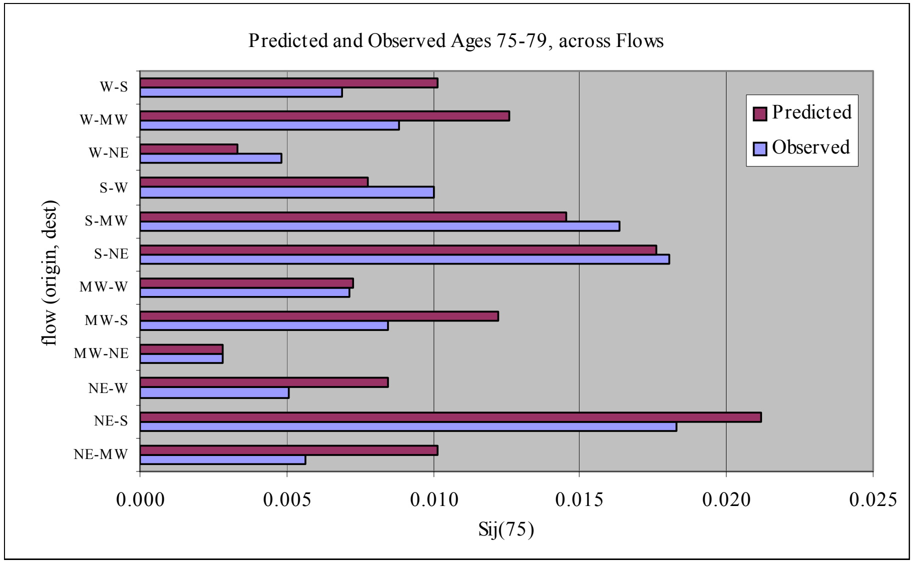


Assessment of Predicted Values, 1995-2000

R-squared Values by Age

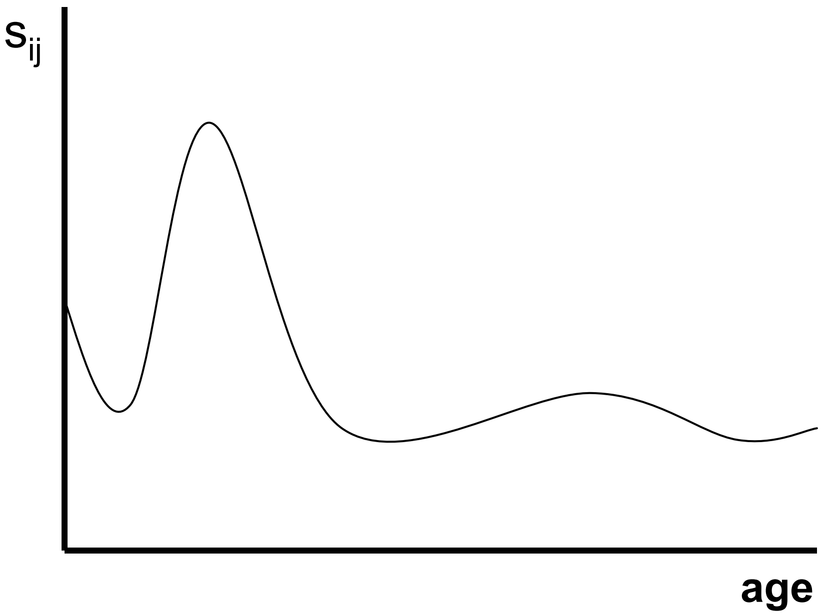
Age	R-sqr
0-4	0.976
5-9	0.939
10-14	0.882
15-19	0.884
20-24	0.969
25-29	0.970
30-34	0.895
35-39	0.896
40-44	0.951
45-49	0.924
50-54	0.960
55-59	0.977
60-64	0.971
65-69	0.906
70-74	0.738
75-79	0.792





POSSIBLE COMPLICATIONS:

- Model Migration Schedule with Retirement Peak



1985-1990: Northeast to South, Midwest to South, Midwest to West

- Violation of Temporal Assumptions

POSSIBLE SOLUTIONS:

If there *are* retirement peaks:

- Regression analysis, where retirement peak flows are estimated separately from non-retirement peak flows
- Improved OLS by estimation:

Sij(x) = *a*+ *b(x)*S_{ij}(-5) + *c(x)*$_{i}$$K_{j}$(+)% + *error term*

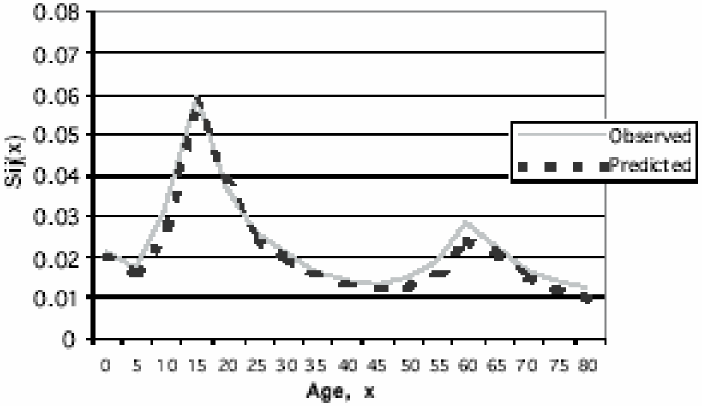
- Logged-odds estimation
- Include more time periods: Geographical Analysis, 30(1):2004 examines 1960-1990 data

If there are retirement peaks in the previous periods, but not in the current time period:

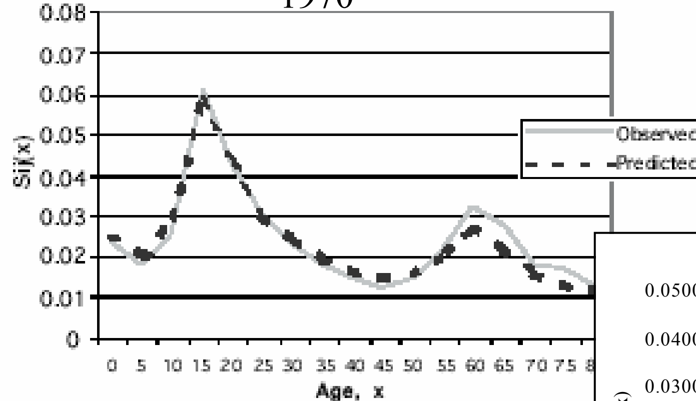
- Use Age-to-Infant Ratios from more similar flows, perhaps find other contextual information that would help predict discrepancies in the age profile of migration

Retirement Peak found through time, 1960-1990:

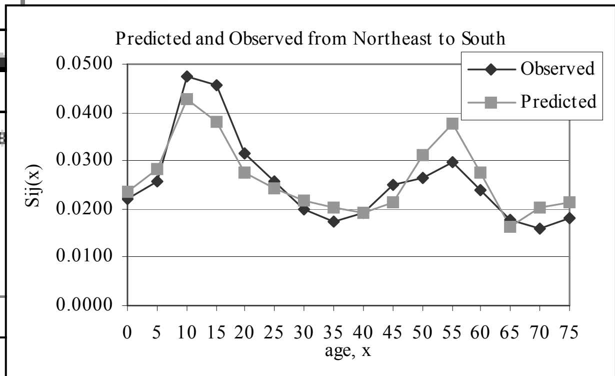
1960



1970

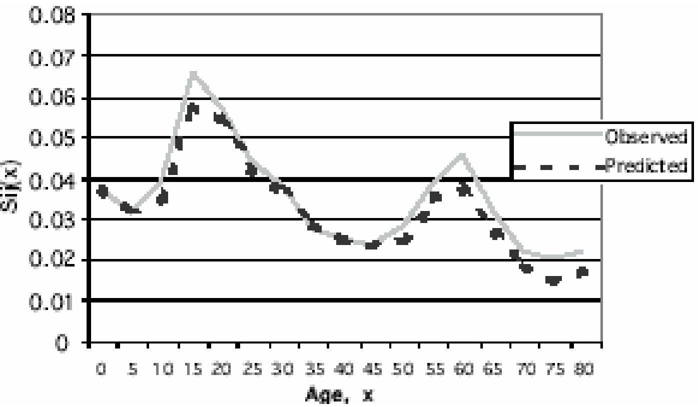


2000

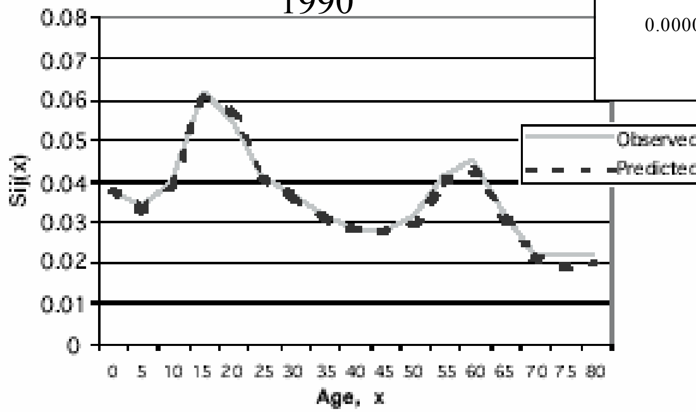


ATI Ratio

1980



1990



Predicted and Observed Migration Schedules from Northeast to South

Issues for Future Research

- Using the birthplace-specific population stocks of infants, other age-specific migration flows can be inferred.

- How well does this methods work for other countries?
- Can excluding certain outliers in the data, by using confidence intervals, help improve this method?