Estimating Complete Migration Probabilities from Grouped Data: A Methods Protocol for a Human Internal Migration Database

Sigurd Dyrting[1] and Andrew Taylor[2]

1 Extended Abstract

1.1 Topic

Internal migration intensities, along with fertility and mortality rates, are the primary processes used to understand the dynamics and spatial structure of population change. The development of a theoretical framework of migration began with Ravenstein's observation that human movement was patterned (Ravenstein, 1885) and evolved to a characterization of its diffusive nature as a balance between new opportunities at the horizon and intervening ones (Stouffer, 1940), the impetus being driven by differences in actual and perceived positive and negative factors at the origin and destination and frustrated by intervening obstacles (Lee, 1966), the mechanism part of a mobility transition progressing sequentially in time and radiating in space fueled by and interacting with the vital transition and more generally a community's passage through modernization (Zelinsky, 1971).

1.2 Theoretical Focus

Incorporating migration into demography's quantitative framework allows a description of population change in size and composition across both time and space (Rogers, 1975). The fundamental variable is a complete schedule of migration rates, a matrix of probabilities that a person at place O with age x will be at place D with age $x + n$ an interval n years later, where O and D are origin and destination labels of a given spatial decomposition and age x ranges by single year increments from birth to the limit of human life ω. As well as allowing the fore mentioned calculations at high age resolution, robust methods for estimating complete migration probability schedules are a necessary precursor for attacking one of the outstanding problems frustrating spatial demography: data. In migration studies, mathematical and conceptual frameworks are more advanced than data, in particular for the

[1] Northern Institute, Charles Darwin University, Darwin, Northern Territory, Australia, sigurd.dyrting@cdu.edu.au

[2] Northern Institute, Charles Darwin University, Darwin, Northern Territory, Australia, andrew.taylor@cdu.edu.au

vital processes there exist public repositories of historical data harmonized to a common format and spanning a range of countries (HFD, 2023; HMD, 2023), but for internal migration there is none.

1.3 Methods

A solution to the data problem in spatial demography will require a method for inferring migration probabilities from age-specific origin-destination matrices that can be applied to both single-year and age-grouped data. While current methods based on splines and model migration schedules (Rogers et al., 2010) have their advantages, they both have limitations that make them unsuitable as a general framework for estimating migration probabilities from grouped data. Spline methods are easy to calibrate and have great flexibility in the age profiles they can fit because they only assume the profile is locally polynomial, however they make assumptions about the population distribution or migration probability over each age group which are not likely to hold as the length of the group interval increases. Model migration schedules have the advantage that they will, when properly calibrated, produce complete schedules that reflect the paradigm their parameters encode, but their accuracy and fidelity are constrained by the assumed functional form.

Recently, the P-TOPALS method has been developed to smooth sample out-migration probabilities (Dyrting, 2020), and Dyrting & Taylor (2021) have shown how it can be combined with a P-spline approach for smoothing destination-specific migration ratios to estimate origin-destination-specific migration probabilities for single year of age data. In this paper we extend both P-TOPALS and P-spline approaches to the case of grouped ages with the aim of developing a general framework for estimating migration probabilities from age-specific origin-destination matrices that combines the strengths of both splines and model migration schedules (flexibility, ease of calibration, ability to specify views on the reasonable form of the age distribution), accounts for sample noise, is stable when the number of age intervals becomes small, and which can be applied to a general age abridgement structure.

1.4 Data

There are 149 microdata samples in IPUMS International (Minnesota Population Center, 2020) with variables that allow the calculation of sample migration probabilities between first-level administrative units. Of these, 76 samples are age-grouped or require grouping because their Whipple index (Whipple, 1919) indicated significant age-heaping. For samples with age-heaping, age was aggregated into five-year age groups. These samples were then examined for preferences in reporting age with end digit 0 over end digit 5. If a preference was evident visually or if the sawtooth index (Riffe et al., 2019) was high, ages over a given value (usually 40 but for some samples lower) were aggregated into ten-year groups. Samples with an open age group were examined for age overstatement, and if detected the open age interval was not included.

1.5 Expected Findings

We have already completed estimation using our method and have begun comparing it with current approaches. We expected to show that our method is a practical tool for estimating migration probabilities across a wide range of data quality conditions.

2 References

Dyrting, S. (2020). Smoothing migration intensities with P-TOPALS. *Demographic Research*, *43*(55), 1607–1650. https://doi.org/10.4054/DemRes.2020.43.55

Dyrting, S., & Taylor, A. (2021). Smoothing destination-specific migration flows. *Annals of Regional Science*, *67*, 359–383. https://doi.org/10.1007/s00168-021-01051-4

HFD. (2023). *Human Fertility Database* [Data set]. Max Planck Institute for Demographic Research (Germany); Vienna Institute of Demography (Austria). Available at https://www.humanfertility.org.

HMD. (2023). *Human Mortality Database* [Data set]. Max Planck Institute for Demographic Research (Germany); University of California, Berkeley (USA); French Institute for Demographic Studies (France). Available at https://www.mortality.org.

Lee, E. S. (1966). A theory of migration. *Demography*, *3*(1), 47–57. https://doi.org/10.2307/2060063

Minnesota Population Center. (2020). *Integrated public use microdata series, international: Version 7.3* [Data set]. https://doi.org/10.18128/D020.V7.3

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, *48*(2), 167–235. https://doi.org/10.2307/2979181

Riffe, T., Aburto, J., Alexander, M., Fennell, S., Kashnitsky, I., Pascariu, M., & Gerland, P. (2019). *DemoTools: An R package of tools for aggregate demographic analysis*.

Rogers, A. (1975). *Introduction to multiregional mathematical demography*. John Wiley & Sons.

Rogers, A., Little, J., & Raymer, J. (2010). *The indirect estimation of migration* (1st ed.). Springer.

Stouffer, S. A. (1940). Intervening opportunities: A theory relating mobility and distance. *American Sociological Review*, *5*(6), 845–867. https://doi.org/10.2307/2084520

Whipple, G. C. (1919). *Vital statistics: An introduction to the science of demography*. John Wiley & Sons.

Zelinsky, W. (1971). The hypothesis of the mobility transition. *Geographical Review*, *61*(2), 219–249. https://doi.org/10.2307/213996