Modeling age patterns of migration in small areas: a flexible framework to account for unexpected deviations

Jessie Yeung* Monica Alexander†

Extended abstract submitted to PAA 2023

Abstract

We propose a general model framework to estimate age-specific migration rates at the subnational level. The model can be decomposed into an expected level — which consists of an overall mean age schedule plus a local area level — and deviations away from expected, which are smoothed over age and time. This modeling framework allows for reliable estimates of local-area migration by age to be made, and also allows deviations away from the expected level to be understood from a temporal perspective. We show preliminary results as applied to in-migration rates to SA2 levels in Australia. Future work will focus on using this framework to understand how local-area migration rates have been lower (or higher) than expected since the onset of the Covid-19 pandemic.

1 Introduction

Reliable estimates of local-area migration rates by age are an important input to population projection models, used by local governments for policy planning and resource allocation. This is particularly pertinent in recent times given the likely substantial impact of the Covid-19 pandemic on internal migration rates. Good estimates and projections of likely migration patterns are needed in order for policymakers to fully understand the impact of the pandemic on migration both in the short and long term.

However, migration rates are often difficult to estimate across small areas, due to small population sizes that make the raw data erratic and underlying trends unclear. From a modeling perspective, obtaining reliable estimates of age-specific mortality rates is particularly challenging because age patterns are inherently non-linear. While the peak age-specific migration rates often occurs at around the young working ages, there are usually secondary retirement and post-retirement peaks

*Statistics Canada. jessie.yeung@mail.utoronto.ca
†University of Toronto. monica.alexander@utoronto.ca

at older ages (Rogers & Castro, 1981). In order to try and capture these multiple peaks, Rogers and Castro (1981) formulated a multi-exponential model, which in later extensions requires up to 13 parameters to be estimated (Rogers & Little, 1994). In practice, estimation of so many parameters, many of which are highly correlated, is difficult, particularly if data are noisy.

In this paper we propose a general model framework to estimate age-specific migration rates at the subnational level. The model can be decomposed into an 'expected level' — which consists of an overall mean age schedule plus a local area level — and deviations away from expected, which are smoothed over age and time. This modeling framework allows for reliable estimates of local-area migration by age to be made, and also allows deviations away from the expected level to be understood from a temporal perspective. We show preliminary results as applied to in-migration rates to SA2 levels in Australia. Future work will focus on using this framework to understand how local-area migration rates have been lower (or higher) than expected since the onset of the Covid-19 pandemic.

2 Modeling Framework

Define $y_{a,r}$ to be the number of persons of age a migrating in (or out) to region r. Define $P_{a,r}$ to be the population aged a in region r. We assume migrant counts are Poisson distributed as follows:

$$y_{a,r} \sim \text{Poisson}(\mu_{a,r} \cdot P_{a,r})$$

Our goal is to estimate migration rates $\mu_{r,a}$. We model these rates on the log scale with the general form

$$\log \mu_{r,a} = \tau_a + \gamma_r + \delta_{a,r}$$

Where

- τ_a is the mean migration age schedule across all regions
- γ_r is a region-level intercept, which represents migration rates that are higher or lower than average
- $\delta_{a,r}$ are age-specific deviation in a particular region

We think of $\tau_a + \gamma_r$ as the 'expected' migration age schedule in region r and $\delta_{a,r}$ to be deviations away from that expected level. In practice, the $\delta_{a,r}$ need to be constrained and smooth to ensure identifiability of the model, and also to obtain reasonable age-specific migration curve estimates. We discuss two options for modeling set-ups below, and how they relate to previous work.

2.1 Modeling deviations with a random walk

A first model set-up assumes that the deviations $\delta_{a,r}$ can be modeled with a random walk over age:

$$\delta_{a,r} \sim N(\delta_{a-1,r}, \sigma_\delta^2)$$

This set-up assumes that the deviation away from the expected level at a particular age a is similar to the value of the deviation observed in the previous age group. In practice, this model set-up smooths deviations over age, with the smoothness depending on the value of the variance σ_δ^2. To ensure identifiability, we constrain the sum of all $\delta_{a,r}$ across all regions to be zero.

We estimate parameters in a Bayesian framework, which allows us to fully consider different types of uncertainty in the model and data. For this modeling strategy, we need to place priors on all parameters. We use weakly-informative standard normal priors:

$$\tau_a \sim N(0, 1)$$

$$\gamma_r \sim N(0, 1)$$

$$\sigma_\delta \sim N^+(0, 1)$$

The full model is then

$$y_{a,r} \sim \text{Poisson}(\mu_{a,r} \cdot P_{a,r})$$

$$\log \mu_{r,a} = \tau_a + \gamma_r + \delta_{a,r}$$

$$\delta_{a,r} \sim N(\delta_{a-1,r}, \sigma_\delta^2)$$

$$\sum_r \delta_{a,r} = 0$$

$$\tau_a \sim N(0, 1)$$

$$\gamma_r \sim N(0, 1)$$

$$\sigma_\delta \sim N^+(0, 1)$$

2.2 Modeling deviations with penalized splines regression

Another option is to model the deviations $\delta_{a,r}$ using penalized splines (P-splines). In particular, P-splines offers a smoothing penalty based on second-order differences that is numerically stable and relatively straighforward to implement (Eilers & Marx, 1996). This method imposes a certain amount of smoothness on the δ's over age:

$$\delta_{r,a} = \sum_k B_{a,k} \alpha_{k,r}$$

$$\alpha_{k,r} \sim N(\alpha_{k-1,r}, \sigma_\alpha^2)$$

where

- $B_{a,k}$ is the cubic splines basis matrix computed from the data and known knot locations
- $\alpha_{k,r}$ is a P-splines parameter that is estimated in the model.

This set-up is related to previous modeling approaches, particularly TOPALS and P-TOPALS. The TOPALS model introduced by de Beer (2011, 2012) can be used to smooth several types of age-specific functions by incorporating a standard age schedule and also applying a linear spline to the ratios between the age-specific rates and the standard age schedule. P-TOPALS is a further extension of the TOPALS method where the deviations from the standard age schedule are modeled using P-splines (Dyrting, 2020). However, our approach differs from P-TOPALS since our model's standard age schedule is estimated instead of inputted in the model as data.

Our P-splines model follows a similar set-up to the random walk model above, including a Poisson likelihood, a mean migration age schedule (τ_a), a region-specific intercept (γ_r), and age-specific deviations for each region ($\delta_{a,r}$). We also constrain the sum of all $\delta_{a,r}$ across all regions to be zero for identifiability.

This yields the following full model:

$$y_{a,r} \sim \text{Poisson}(\mu_{a,r} \cdot P_{a,r})$$

$$\log \mu_{r,a} = \tau_a + \gamma_r + \delta_{a,r}$$

$$\delta_{r,a} = \sum_k B_{a,k} \alpha_{k,r}$$

$$\alpha_{k,r} \sim N(\alpha_{k-1,r}, \sigma_\alpha^2)$$

$$\sum_r \delta_{a,r} = 0$$

$$\tau_a \sim N(0, 1)$$

$$\gamma_r \sim N(0, 1)$$

$$\sigma_\alpha \sim N^+(0, 1)$$

Extension to multiple time periods The previous discussion just considers one time period only. To extend to multiple time periods, let $y_{a,r,t}$ to be the number of persons of age a migrating in (or out) to region r at time t. We again assume migrant counts are Poisson distributed:

$$y_{a,r,t} \sim \text{Poisson}(\mu_{a,r,t} \cdot P_{a,r,t})$$

Now we assume the migration rates on the log scale are modeled as

$$\log \mu_{r,a,t} = \tau_{a,t} + \gamma_r + \delta_{a,r,t}$$

To model the deviations away from the expected level over time, we can now consider extending the models above to also smooth the data over the temporal dimension. For example, we could employ a two-dimensional random walk:

$$\delta_{a,r,t} \sim N(\delta_{a-1,r,t-1}, \sigma_\delta^2)$$

3 Preliminary application to small-area migration in Australia

In this section, we show preliminary results of the models in which deviations are smoothed with a random walk (section 2.1) and with P-splines (section 2.2).

3.1 Data

We fit our models on Statistical Areal Level 2 (SA2) migration data from the 2016 Australian census, obtained from the Australian Bureau of Statistics. In particular, we have migration matrices which provide migration counts for each destination and place of usual residence 5 years prior, for both sexes and each 5-year age group. From these matrices, we are able to obtain age-specific 5-year migration rates for each SA2 region in Australia. This yields age-specific migration curves for each of the 2,310 SA2 regions that we wish to smooth using our model.

3.2 Preliminary results

To illustrate the models, we run them on in-migration data for males, across all SA2 regions in Tasmania, Australia. This generates a smoothed migration curves for each of the 99 SA2 regions in Tasmania.

Figure 1 shows the standard age schedule estimated from all SA2 regions in Tasmania. The smoothed migration curves for the SA2 regions of Derwent Park - Lutana and Sandy Bay are shown in Figure 2. It shows the data provided by ABS (in black), the estimated standard age schedule (same as figure 1, in blue), and the estimated fitted curve for a particular region (in red).

Figures 3 and 4 show similar results of the standard age schedule and smoothed migration curves for Derwent Park - Lutana and Sandy Bay when the deviations (δ's) are smoothed with P-splines.

Both the Random Walk and P-splines models yield similar results although the added complexity of the P-splines model provides further smoothing for the deviations. Since our models use a Poisson likelihood, the fitted migration curves follow the data more closely when the population is larger.

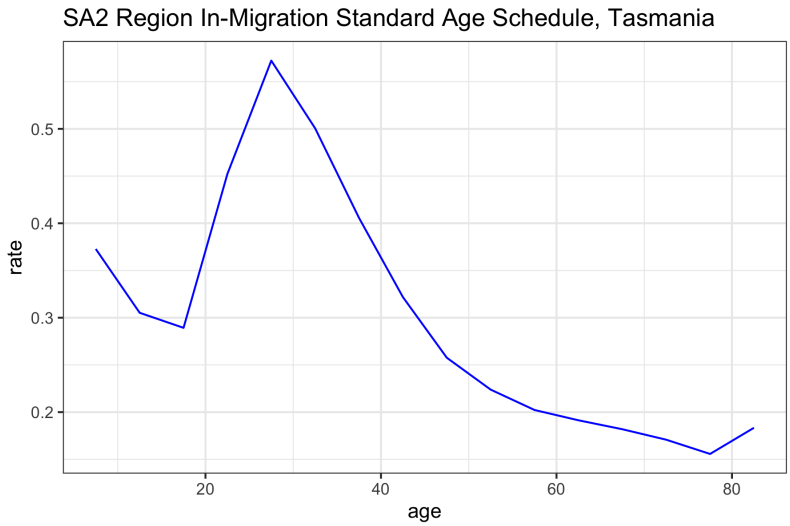

Figure 1: Smoothed in-migration age schedule for males in Tasmania from the Random Walk model.

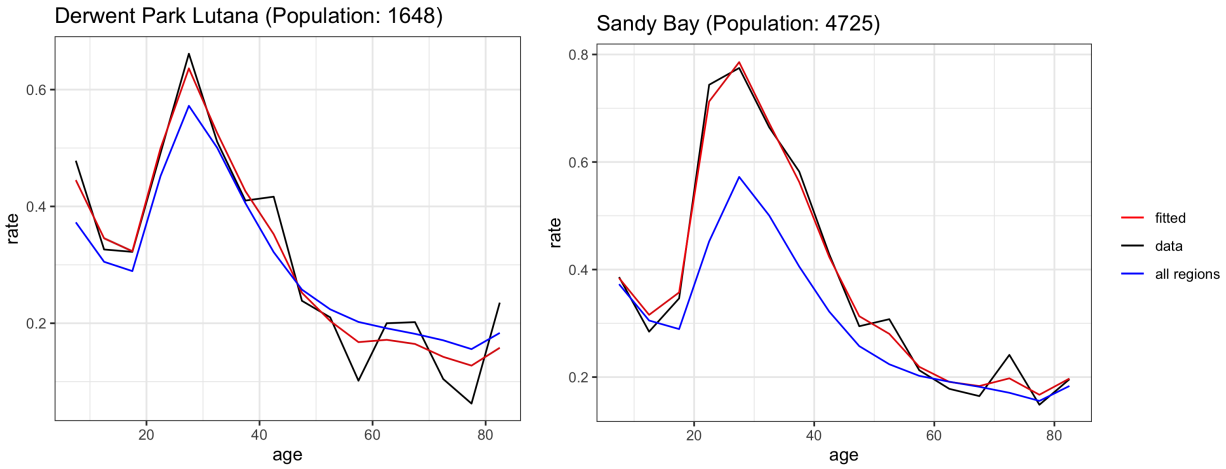

Figure 2: Smoothed in-migration patterns for males in two SA2 regions in Tasmania from the Random Walk model.

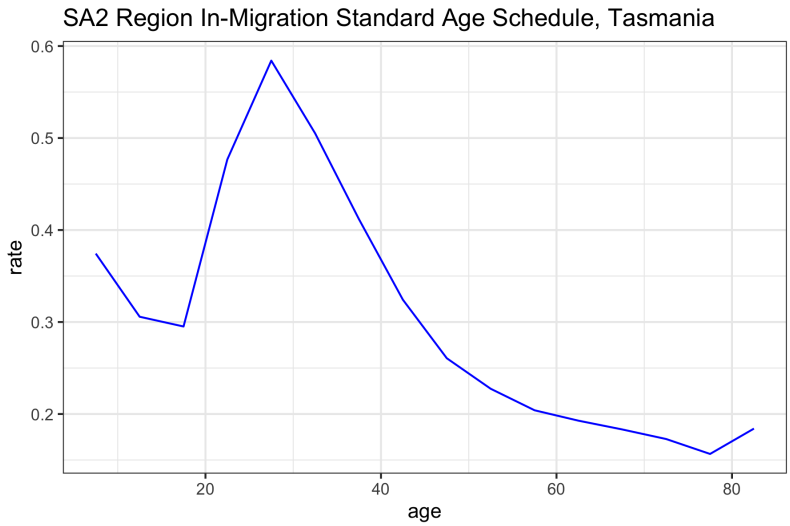

Figure 3: Smoothed in-migration age schedule for males in Tasmania from the P-splines model.

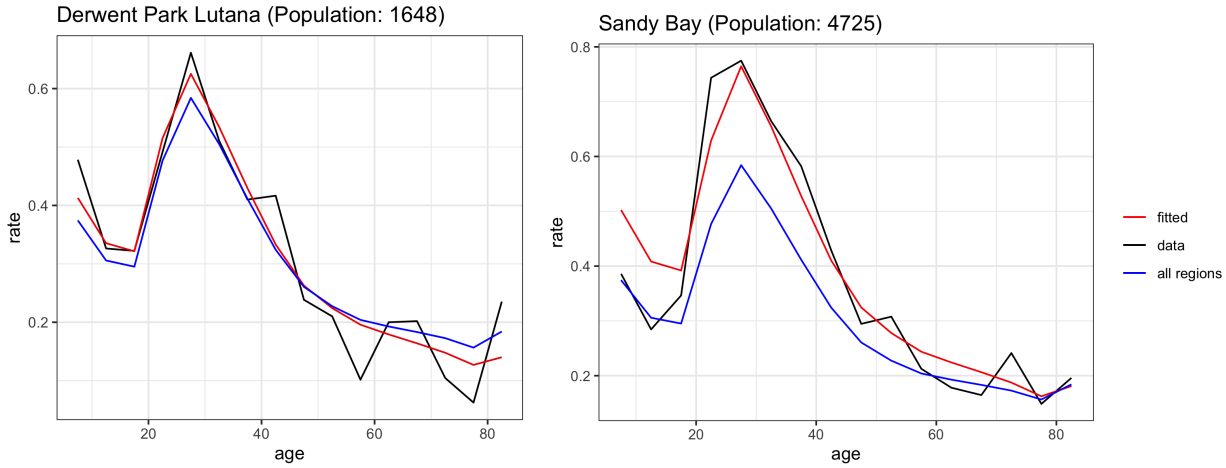

Figure 4: Smoothed in-migration patterns for males in two SA2 regions in Tasmania from the P-splines model.

However, when the population is small, the fitted curve relies more heavily on the position and/or shape of the standard age schedule. We can see this demonstrated when comparing the results for Derwent Park - Lutana which has a smaller population relative to that of Sandy Bay.

4 Summary and future work

In this abstract we proposed a general modeling framework to estimate age-specific migration rates at the subnational level. Our proposed framework consists of an expected level, which is a global mean plus area-level intercept, and deviations, which are smoothed over age and time. Preliminary results fitted to SA2 level data in Australia in one year show promising results. Future work will focus on estimation over time, using data from Australia and the United States that covers a period before and after the Covid-19 pandemic. We envisage estimates will help to understand the impact of the pandemic on internal migration, and how migration to and from subnational regions has rebounded (or otherwise) since 2020.

References

deBeer, J. (2011). A new relational method for smoothing and projecting age-specific fertility rates: TOPALS. *Demographic Research*, *24*, 409–454.

deBeer, J. (2012). Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research*, *27*, 543–592.

Dyrting, S. (2020). Smoothing migration intensities with p-TOPALS. *Demographic Research*, *43*, 1607–1650.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*(2), 89–121.

Rogers, A., & Castro, L. J. (1981). *Model migration schedules.*

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