Bayesian implementation of Rogers-Castro model migration schedules: An alternative technique for parameter estimation

Jessie Yeung Monica Alexander Tim Riffe

Abstract

BACKGROUND

The Rogers-Castro model migration schedule is a key model for migration trends over the life course. It is applied in a wide variety of settings by demographers to examine the relationship between age and migration intensity. This model is nonlinear and it can have up to 13 parameters, which can make estimation difficult. Existing techniques for parameter estimation can lead to issues such as nonconvergence, sensitivity to initial values, or optimization algorithms that do not reach the global optimum.

OBJECTIVE

We propose a new method of estimating the Rogers-Castro model migration schedule parameters that overcomes most common difficulties.

METHODS

We apply a Bayesian framework for fitting the Rogers-Castro model. We also provide the R package 'rcbayes' with functions to easily apply our proposed methodology.

RESULTS

We illustrate how this model and R package can be used in a variety of settings by applying it to data from the American Community Survey.

CONTRIBUTION

We provide a novel and easy-to-use approach for estimating Rogers-Castro model parameters. Our approach is formalized in an R package which makes parameter estimation and Bayesian methods more accessible for demographers and other researchers.

1. Introduction

One of the most important considerations for demographers dealing with migration patterns is the fundamental relationship between migration and age. This relationship parallels how migration is influenced by multiple transitions over the life course such as starting or finishing education, new employment opportunities, or retirement (Preston, Heuveline, and Guillot 2000). These regularities in migration intensity over the life course allow demographers and policymakers to model migration by age. This fundamental relationship between migration and age is most notably described by Rogers and Castro who were the first to introduce a mathematical model for migration in the form of a flexible multiexponential migration model (Rogers, Racquillet, and Castro 1978; Rogers and Castro 1981). Initial versions of the model included families of 7-parameters, 9-parameters, and 11-parameters, but was later extended to include the 13-parameter schedule (Rogers and Little 1994). Over the past forty years this model has become well-known among demographers as the *Rogers-Castro model migration schedule* or the *multiexponential model schedule*. Rogers and Castro argues in a variety of works that this model is well-suited to describe the migration age schedules that appear across regions of various sizes and across gender and ethnic subgroups (Rogers and Castro 1981).

The standard Rogers-Castro migration age schedule is shown in Figure 1. Migration intensities typically peak in early adulthood due to the relatively higher number of transitions involving education, employment, and partner formation. Around retirement age there is sometimes another smaller peak resulting from the transition out of the workforce as individuals form new living arrangements for retirement. Among the oldest age groups migration intensities may gradually increase as individuals move into institutions or other homes that provide additional care and support. Finally, the migration intensity of children tends to mirror that of their parents. For example, migration intensity is high in the youngest ages as parents experience the high migration intensities of early adulthood.

Figure 1: Standard shape of Rogers-Castro migration age schedule

Note: This figure shows the standard shape of the Rogers-Castro migration age schedule with labels for the four main components (preworking age, working age, retirement age, and postretirement age) that each correspond with migration-related lifespan stages. These four components are further discussed in later sections.

Age-migration models are useful in practice because they smooth noisy data, estimate missing data, allow for prediction of future trends, and help to model other demographic trends that rely on migration. In particular, the Rogers-Castro model migration schedule has successfully been applied to interregional migration in North America (Rogers and Little 1994; Raymer and Rogers 2007; Raymer and Rogers 2008), the United Kingdom (Bates and Bracken 1982), within Japan, Korea and Thailand (Kawabe et al. 1990), and South Africa (Hofmeyer 1988). The model has also been used to inform population projects in the case of Canada (George 1994). Some demographers argue for adjustments to the standard Rogers-Castro model in order to accommodate for idiosyncratic features found in specific populations, such as a second retirement peak (Warnes 1992) or a student peak (Wilson 2010). However, the main components of the standard Rogers-Castro model are able to capture age-based migration trends in the majority of cases, and the model is used in many production settings such as at the World Population Prospects (United Nations 2019).

There are a number of existing ways of estimating and fitting the Rogers-Castro migration schedule. One approach is to use hand calculations, whereby with the appropriate formulas one can estimate the model parameters with a simple calculator (Rogers, Castro, and Lea 2005). Other approaches involve using various software based optimization procedures such as *Solver* in Excel (Little and Dorrington 2013), Table Curve 2D (Rogers and Raymer 1999), *nlminb* in R (Ruiz-Santacruz 2021), Gauss-Newton algorithm (Little and Dorrington 2013), and several others. In practice, the Rogers-Castro model can be hard to fit with standard optimization algorithms due to its nonlinearity and large number of parameters. Additionally, the results are highly sensitive to initial conditions and there is never a guarantee that the global minimum has been reached.

In this paper, we propose an alternative method of estimating Rogers-Castro model parameters in a Bayesian framework. Implementing the Rogers-Castro framework in a Bayesian model has been mentioned sparingly in the literature (Congdon 2008), although no current tools exist that implement this methodology in a manner that is easy and simple to use. Our work is formalized in the R package *rcbayes*, which is available on CRAN. This package allows users to obtain Bayesian estimates of Rogers-Castro parameters and is user-friendly for those less familiar with the specifics of running Bayesian models. We will demonstrate in this paper how to use the *rcbayes* package and how this methodology mitigates issues present in other methods of estimation.

2. Rogers-Castro model migration schedule

The Rogers-Castro model migration schedule consists of four components to capture trends in each of the preworking ages, working ages, retirement ages and postretirement ages. These four components were previously visualized in Figure 1. The model is flexible in that any of the four components can be included or excluded based on the data and situation. Assuming that all four components are included, the full mathematical specification for the migration rate at age x is:

$$
\begin{aligned}
m(x) = \; & a_1 \exp\{-\alpha_1 x\} + && \text{(preworking)} \\
& a_2 \exp\{-\alpha_2(x-\mu_2) - \exp[-\lambda_2(x-\mu_2)]\} + && \text{(working)} \\
& a_3 \exp\{-\alpha_3(x-\mu_3) - \exp[-\lambda_3(x-\mu_3)]\} + && \text{(retirement)} \\
& a_4 \exp\{\alpha_4 x\} + && \text{(postretirement)} \\
& c && \text{(overall)}
\end{aligned}
$$

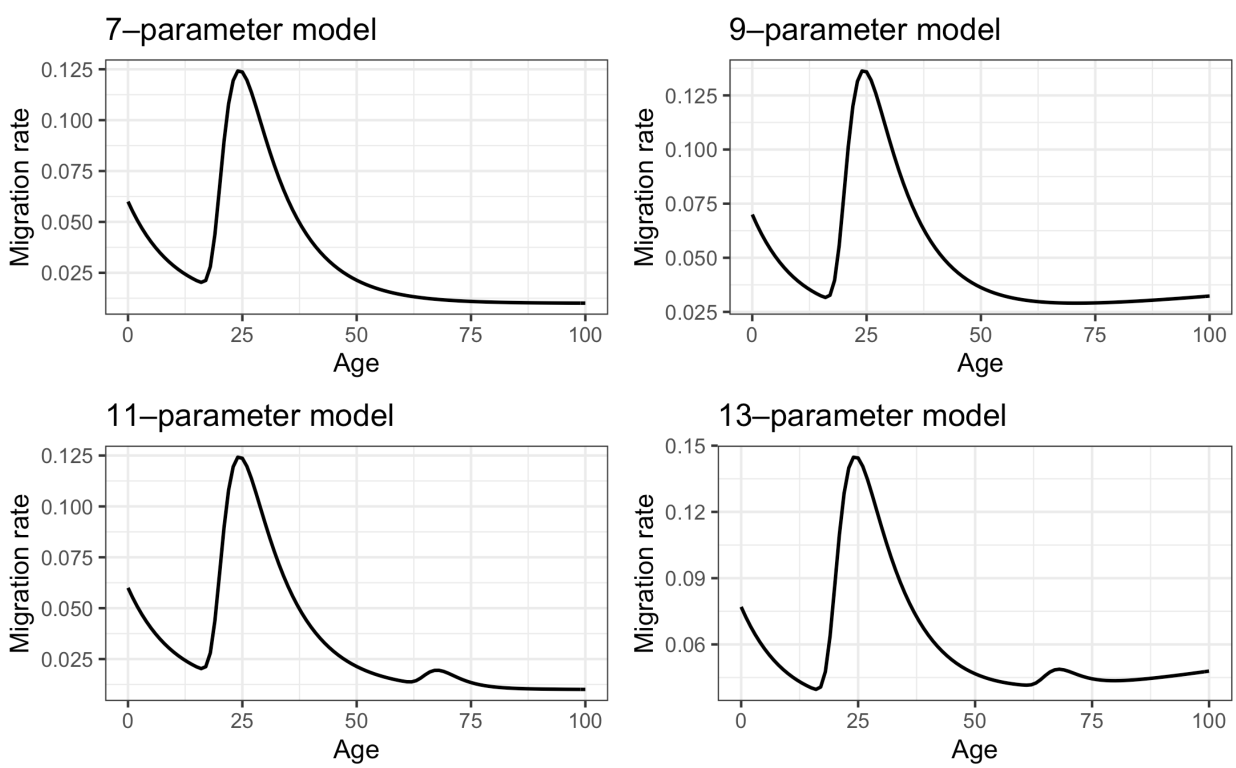
Each term represents one of the four components whereas the final c parameter is related to the baseline level of migration. The parameters within each term captures the intensity or shape of migration in each stage. In particular, the $\lambda_k$parameters influence the steepness of the ascending side of the peaks and the a_k parameters influence the steepness of the descending side. Together the $\alpha_k$and $\lambda_k$parameters capture the rate of change over age, thereby impacting the overall shape of that component. Finally, the μ_2 and μ_3 parameters influence the age for the working age peak and retirement peak, respectively.

This migration model has up to 13 parameters if all four components are included. In practice, there are four combinations of components that are most commonly seen and used (Rogers, Little, and Raymer 2010):

- 7 parameter model including preworking and working age components
- 9 parameter model including preworking, working, and postretirement age components
- 11 parameter model including the preworking, working, and retirement age components
- 13 parameter model that includes all four components

Figure 2 shows the shapes for each of these commonly found migration schedules.

Figure 2: Most Common shapes for Rogers-Castro migration age schedules



3. Previous approaches

Most approaches to estimate Rogers-Castro parameters involve optimization algorithms. One common approach is to use *Solver* in Excel. This method involves first interpolating one-year migration intensities using a procedure such as cubic splines, in the case that the data does not come in one-year age intervals (e.g. five-year migration intensities are common) (Little and Dorrington 2013). Additionally, the user manually sets initial estimates as starting values, and results can be highly sensitive to these chosen values. Since there is always the risk that the output is not a global optimum, it is recommended to choose starting values based on known parameters of a "similar" curve. However, results of a similar curve are not always going to be available, particularly in unique cases. Nevertheless, convergence is still difficult to achieve in the 11-parameter or 13-parameter models. All of this suggests that it will often take experimentation and several attempts of fine-tuning to achieve acceptable results (Rogers, Castro, and Lea 2005).

Another approach is to use deterministic mathematical formulas that reduce parameter estimation to simple hand calculations. In particular, Rogers, Castro, and Lea provide 3 linear methods of estimating model parameters in this way and argued that these methods perform satisfactorily (Rogers, Castro, and Lea 2005). The main advantage of this method is that it overcomes the issues of non-convergence and instability that inevitably arise with the nonlinear optimization procedures. However, as there are several sets of formulas that one can use which all lead to different results, there is no obvious way of choosing a particular approach. Each approach will be biased in a different way and choosing any approach requires an implicit bias-variance trade-off.

Additionally, it is also possible for these two approaches to be combined. In the case of an augmented version of the Rogers-Castro model that includes a student peak, Wilson presents an approach to estimate those parameters using steps that utilize both Excel *Solver* and formula-based calculations (Wilson 2010). Calculations were used for parameters that are easier to estimate in that way, and Excel *Solver* is used multiple times for a small number of parameters to avoid nonconvergence issues.

Finally, with the increase in popularity of statistical software such as R, Stata, SPSS, and SAS, it only makes sense for the aforementioned estimation approaches to be replicated using this software, particularly as the number of people who are proficient in Excel *Solver* or Table Curve 2D become more limited. Oftentimes maximum likelihood combined with an optimization algorithm can be used to estimate parameters. For example, *migraR* is one particular R package that focuses on estimation of Rogers-Castro model parameters using a gradient decent algorithm (Ruiz-Santacruz 2021). However, traditional linear optimizers may not behave as desired in high-dimension as they are subject to many of the aforementioned pitfalls like sensitivity to initial values and issues reaching the global optimum.

4. Our model and recommended steps

We propose two ways to express the Rogers-Castro model migration schedule in a Bayesian framework. First of all, we offer a Poisson model where the number of (in or out) migrants is modelled with a Poisson distribution where population is used as an exposure. Fitting this model would require data on specified ages or age groups (x_i), the number of age-specific observed migrants (y_i) and age-specific exposures (O_i) such as population or sample sizes. Note that i is the index for age groups where, for example, $i = 1$ represents the first age group. The model is as follows:

$$y_i \sim Poisson(m_i \times O_i)$$

$$\begin{aligned} m_i = {} & a_1 \exp\{\alpha_1 x_i\} \\ & + a_2 \exp\{-\alpha_2(x_i - \mu_2) - \exp[-\lambda_2(x_i - \mu_2)]\} \\ & + a_3 \exp\{-\alpha_3(x_i - \mu_3) - \exp[-\lambda_3(x_i - \mu_3)]\} \\ & + a_4 \exp\{\alpha_4 x_i\} \\ & + c \end{aligned}$$

$$\alpha_1 \sim N(0, 1)$$
$$a_1 \sim N(0, 0.1)$$

$$\alpha_2 \sim N(0, 1)$$
$$a_2 \sim N(0, 0.1)$$
$$\mu_2 \sim N(25, 1)$$
$$\lambda_2 \sim N(0, 1)$$

$$\alpha_3 \sim N(0, 1)$$
$$a_3 \sim N(0, 0.1)$$
$$\mu_3 \sim N(65, 1)$$
$$\lambda_3 \sim N(0, 1)$$

$$a_4 \sim N(0, 0.05)$$
$$\lambda_4 \sim N(0, 0.01)$$

$$c \sim N\left(\min_{\mathrm{i}} \{y_i/O_i\}, 0.1\right)$$

Note that in this notation, m_i represents the true age-specific migration rate which is to be estimated in the model.

Second, we propose a Normal model where the migration rate is modelled with a Normal distribution. Fitting this model would require data on specified ages or age groups (x_i) and age-specific observed migration rates (r_i). Additionally, the standard deviation in the likelihood (σ) can be specified as well. If it is not specified, σ will be estimated in the model. The model is as follows:

$$r_i \sim N(m_i, \sigma^2)$$

$$\begin{aligned} m_i = {} & a_1 \exp\{\alpha_1 x_i\} \\ & + a_2 \exp\{-\alpha_2(x_i - \mu_2) - \exp[-\lambda_2(x_i - \mu_2)]\} \end{aligned}$$

$$
\begin{aligned}
&+ a_3 \exp\{-\alpha_3(x_i - \mu_3) - \exp[-\lambda_3(x_i - \mu_3)]\} \\
&+ a_4 \exp\{\alpha_4 x_i\} \\
&+ c
\end{aligned}
$$

$$
\begin{gathered}
\alpha_1 \sim N(0,1) \\
a_1 \sim N(0,0.1)
\end{gathered}
$$

$$
\begin{gathered}
\alpha_2 \sim N(0,1) \\
a_2 \sim N(0,0.1) \\
\mu_2 \sim N(25,1) \\
\lambda_2 \sim N(0,1)
\end{gathered}
$$

$$
\begin{gathered}
\alpha_3 \sim N(0,1) \\
a_3 \sim N(0,0.1) \\
\mu_3 \sim N(65,1) \\
\lambda_3 \sim N(0,1)
\end{gathered}
$$

$$
\begin{gathered}
a_4 \sim N(0,0.05) \\
\lambda_4 \sim N(0,0.01) \\
c \sim N\left(\min_{\mathrm{i}} \{r_i\}, 0.1\right)
\end{gathered}
$$

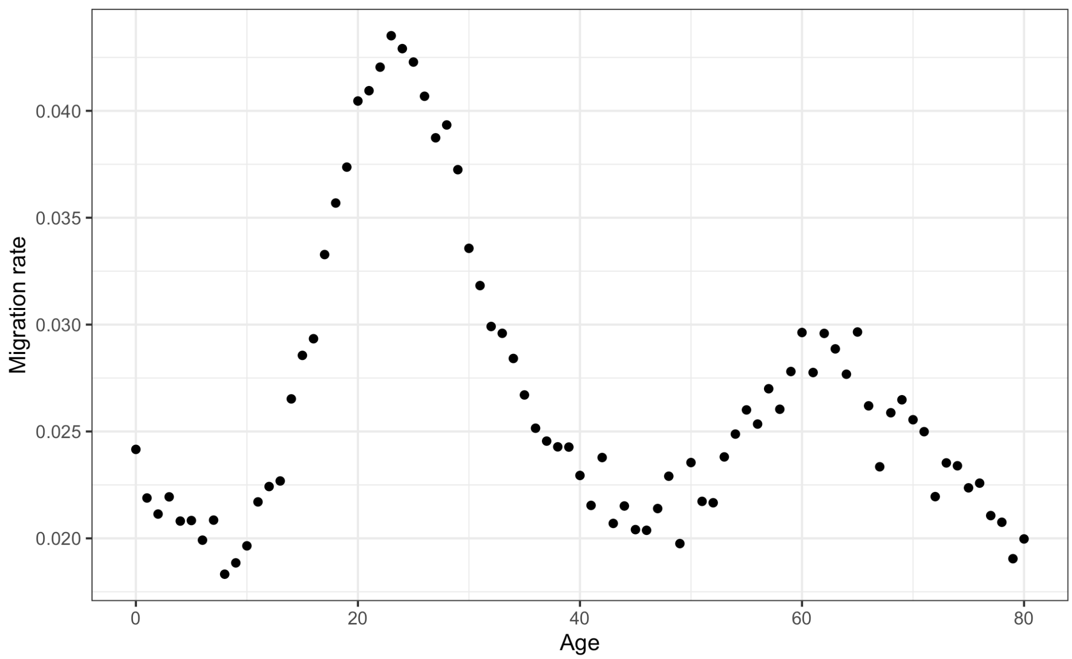
Within the *rcbayes* package, the *mig_estimate_rc* function estimates the parameters of this model using a Markov Chain Monte Carlo (MCMC) algorithm, via the Stan programming language (Carpenter et al. 2017). The Bayesian methods used here set priors for each parameter, which helps with convergence during the estimation process.

The following is a guide of recommended steps for how to use the *rcbayes* package to fit the Rogers-Castro model. The purpose of these steps is to provide users with a general understanding of the recommended workflow. More details on how to use the package functions are provided in package vignettes.

Step 1: Exploratory Data Analysis

As usual, it is recommended to begin analysis by examining numerical and graphical summaries of data, as demonstrated in Figure 3. In this situation, the exploratory data analysis also serves the additional role of helping to decide which components of the Rogers-Castro model to include: preworking age, working age, retirement age, and postretirement age patterns. When fitting the model, we recommend to only include components seen in the data, in order to avoid overfitting or incurring convergence issues.

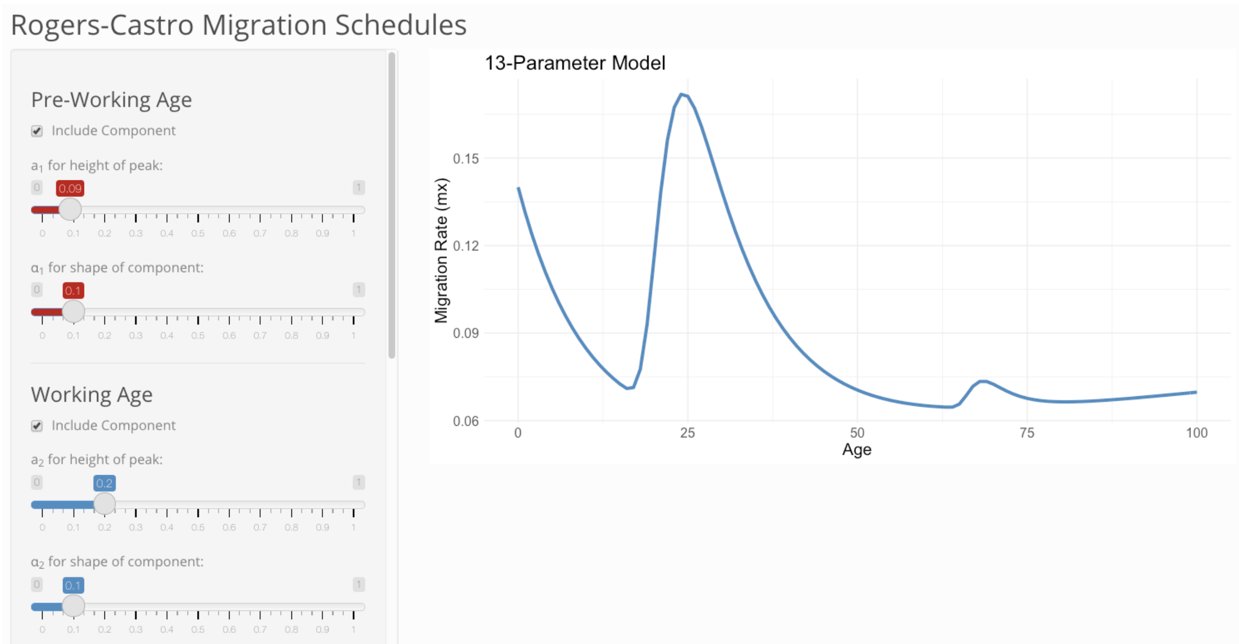
Figure 3: Exploratory visualization for migration rates over the life course



Note: Based on this exploratory analysis, it would be reasonable to fit an 11-parameter model that includes preworking, working, and retirement components only.

To help with this decision we offer an interactive Shiny app in *rcbayes*, as shown in Figure 4. This app gives an easy way to visualize what the multiexponential curve looks like in the presence or absence of any of the 4 components. Note: After loading the package into R, run *rcbayes::interact_rc()* to activate the Shiny app.

Figure 4: Shiny app available through rcbayes for visualizing the multiexponential curve



Step 2: Estimate the model

The next step is to fit the model using the *mig_estimate_rc* function. The choice between fitting the Poisson model or the Normal model will sometimes be guided by the available data. The Poisson model requires (1) a vector of ages, ideally of the age-group midpoints if they span multiple years, (2) a vector of age-specific in-migrant or out-migrant counts, and (3) a vector of age-specific population sizes. The Normal model requires (1) a vector of ages and (2) a vector of age-specific migration rates. For the Normal model, the user can also choose to specify a value for (3) the standard deviation σ across the age-specific migration rates. If this is not specified, the standard deviation will be estimated along with other model parameters. (See Example 1 below for more information on specifying the standard deviation in the Normal model.)

If only rate data is available (as opposed to counts and population sizes) the Normal model is the only option. If migration counts and population counts are available, both the Normal and Poisson models can be fit. The Poisson model will likely be more accurate in its assumption about the underlying random process that generated the data, but the Normal model will sometimes be easier to fit without convergence issues.

The *mig_estimate_rc* function will fit either the Poisson model or the Normal model depending on the arguments provided. In addition, the 4 model components to include should be specified explicitly. Any additional parameters are optional arguments for Stan. The following sample code demonstrates how to use the function.

```
#fit the Poisson model by using ages, migrants, and pop
res <- mig_estimate_rc(ages = age_data, migrants = mig_data, pop = pop_data,
                       pre_working_age = TRUE,
                       working_age = TRUE,
                       retirement = TRUE,
                       post_retirement = FALSE,
                       #optional inputs into stan
                       control = list(adapt_delta = 0.95, max_treedepth = 10)
                       )

#fit the Normal model by using ages and mx (migration rates)
res <- mig_estimate_rc(ages = age_data, mx = mx_data,
                       sigma = 0.002, #optional
                       pre_working_age = TRUE,
                       working_age = TRUE,
                       retirement = TRUE,
                       post_retirement = FALSE,
                       #optional inputs into stan
                       control = list(adapt_delta = 0.95, max_treedepth = 10)
                       )
```

Step 3: Examine Convergence

After running a model with MCMC, it is important to check the convergence of the model. The *check_converge* object in the function output provides model diagnostics. For further guidance on how to verify convergence, the *rcbayes* package includes a vignette “Rogers Castro Migration Models with rcbayes” which explains this step in more detail. In the presence of nonconvergence, there is also another vignette “Achieving Model Convergence With mig_estimate_rc” that explains how this can be resolved. In our experience, the majority of convergence issues can be resolved by setting initial values using the *init_vals* function – especially in the full 13-parameter model.

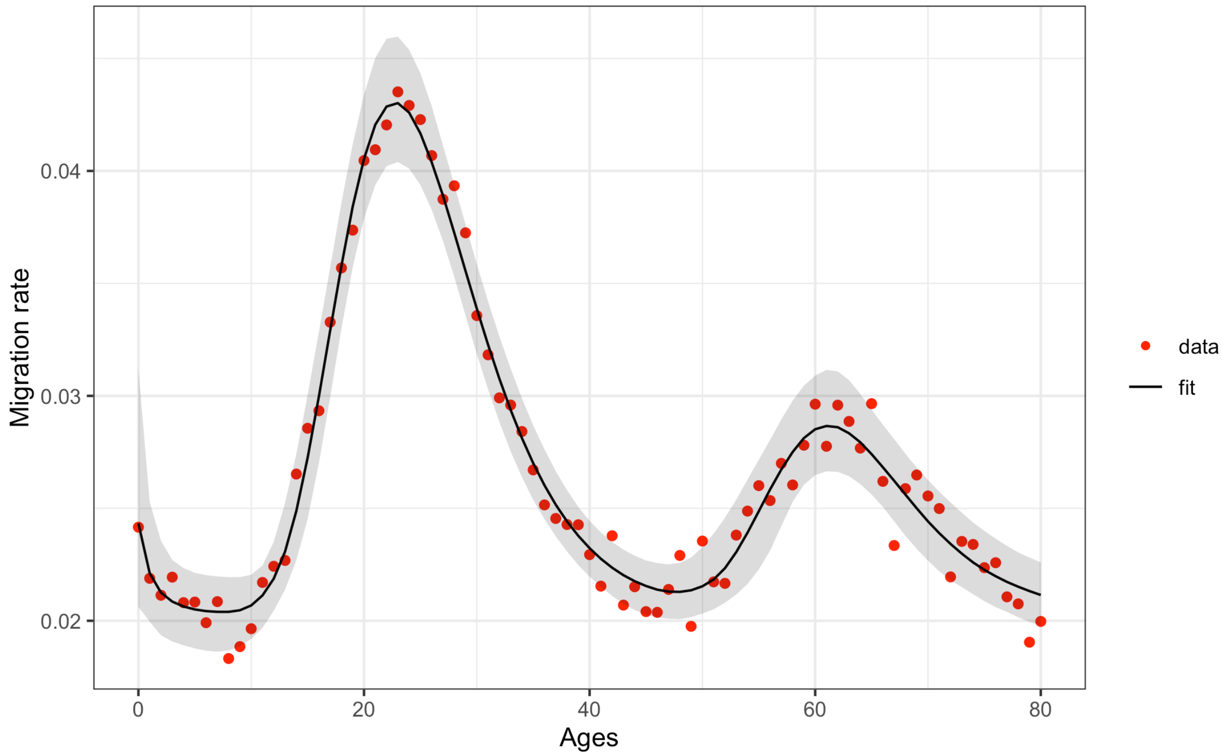
Step 4: Examine Results

The next step is to examine and interpret the fitted model. As previously mentioned, the *check_converge* object in the function output provides model diagnostics. There are two additional objects in the output of *mig_estimate_rc* that provide results. The *pars_df* object shows the median estimate and lower and upper bound of a 95% credible interval for the Rogers-Castro parameters. The *fit_df* object shows the data and estimated median migration rates at each age, along with the lower and upper bound of the 95% credible interval of the fits, and the squared difference between data and the median estimate. A plot of the results can be generated as follows:

```
rc_res[["fit_df"]] %>%
  ggplot(aes(ages, data)) +
  geom_point(aes(color = "data")) +
  geom_line(aes(x = age, y = median, color = "fit")) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  ylab("Migration rate") + xlab("Ages") +
  scale_colour_manual(name = "",
                      values = c(data = "red", fit = "black"),
                      guide = guide_legend(override.aes = list(
                        linetype = c("blank", "solid"),
                        shape = c(16, NA))))
```

This yields a plot similar to the one shown in Figure 5.

Figure 5: Example of result from Bayesian estimation of Rogers-Castro model parameters



Step 5: Sensitivity Analysis of Priors

The final step in the workflow could involve a sensitivity analysis of the prior specification of the model (Depaoli et al., 2020). An important consideration in Bayesian models is that prior specification in a model can impact the posterior estimates of the parameters of interest. Conducting a sensitivity analysis enables researchers to distinguish the effect of choice of priors from the role of data in the resulting estimates. This analysis is conducted by fitting the data on similar models where alternative priors are used, including both informative and non-informative priors. We emphasize that sensitivity analysis results are intended for interpretation purposes and should not alter the final model results. Including these results alongside the original model estimates enhances the discussion and facilitates a better comprehension of the final estimates.

In conducting a sensitivity analysis, the model needs to be altered in ways specific to a particular analysis. When developing the rcbayes package and in particular the *mig_estimate_rc* function, we chose to implement 'weakly informative priors' on all parameters, which are outlined in the previous section. These priors ensure that parameter estimates are within a plausible range of values across a wide range of estimation contexts, and also helps with model convergence. In the current version on CRAN, the rcbayes package does not allow the user to specify priors. However, to aid in the process of sensitivity analysis, we provided a supplementary file is provided with R code that allows changes to the model priors. This code is very similar to the *mig_estimate_rc* function and is intended as a gentle guide for those less familiar with fitting Bayesian models in Stan using R.

5. Sensitivity Analysis

We present a sensitivity analysis of both the Poisson and Normal models by fitting the model on data simulated from a known 11-parameter Rogers-Castro curve. We compare the results from (1) the original model offered in the *rcbayes* package, (2) a model with less informative priors, and (3) a model with very informative priors[1]. Details of this model are provided in Table 1. These models are also fit on several different population sizes, ranging from a total population size of 10,000 to 10,000,000 across all age groups. The results of the sensitivity analysis are shown in Figure 6.

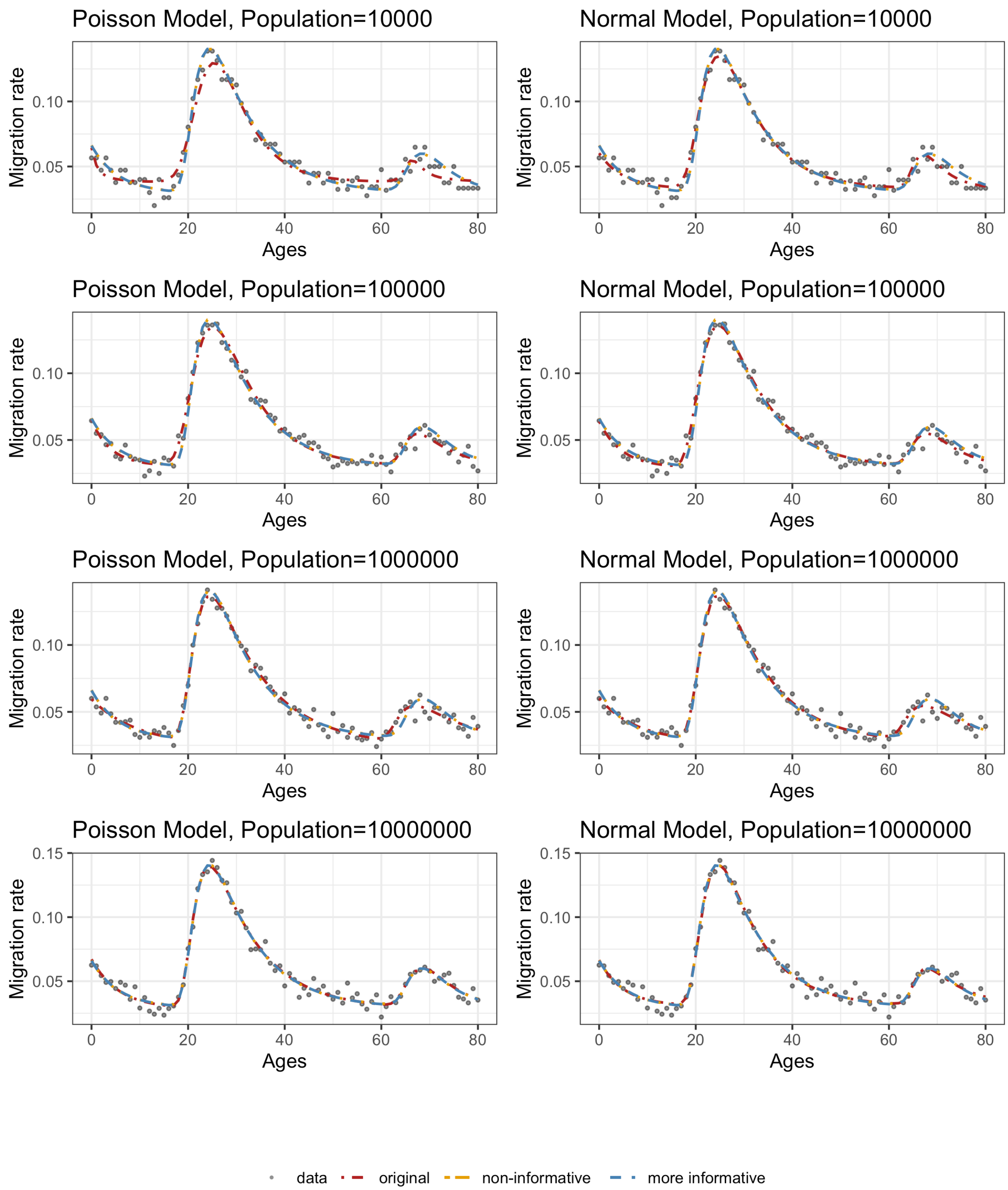
[1] The third model with very informative priors is unrealistically informative, as priors were selected with the known Rogers-Castro curve parameters in mind. This is done to demonstrate the results if the priors reflected extraordinarily accurate prior information of the parameter values.

Table 1: Priors used in sensitivity analysis

Original priors	Less informative priors	More informative priors
$\alpha_1 \sim N(0, 1)$	$\alpha_1 \sim N(0, 10)$	$\alpha_1 \sim N(0.1, 0.1)$
$a_1 \sim N(0, 0.1)$	$a_1 \sim N(0, 10)$	$a_1 \sim N(0.1, 0.1)$
$\alpha_2 \sim N(0, 1)$	$\alpha_2 \sim N(0, 10)$	$\alpha_2 \sim N(0.1, 0.1)$
$a_2 \sim N(0, 0.1)$	$a_2 \sim N(0, 10)$	$a_2 \sim N(0.1, 0.1)$
$\mu_2 \sim N(25, 1)$	$\mu_2 \sim N(1, 100)$	$\mu_2 \sim N(21, 0.5)$
$\lambda_2 \sim N(0, 1)$	$\lambda_2 \sim N(0, 10)$	$\lambda_2 \sim N(0.5, 1)$
$\alpha_3 \sim N(0, 1)$	$\alpha_3 \sim N(0, 10)$	$\alpha_3 \sim N(0.1, 1)$
$a_3 \sim N(0, 0.1)$	$a_3 \sim N(0, 10)$	$a_3 \sim N(0.1, 0.1)$
$\mu_3 \sim N(65, 1)$	$\mu_3 \sim N(1, 100)$	$\mu_3 \sim N(67, 0.5)$
$\lambda_3 \sim N(0, 1)$	$\lambda_3 \sim N(0, 10)$	$\lambda_3 \sim N(0.5, 1)$
$c \sim N\left(\min_{\mathrm{i}} \{r_i\}, 0.1\right)$	$c \sim N(0, 10)$	$c \sim N\left(\min_{\mathrm{i}} \{r_i\}, 0.1\right)$
$\sigma \sim N^+(0, 1)$	$\sigma \sim N(0, 10)$	$\sigma \sim N^+(0, 1)$

Note: Prior on σ only applies to Normal models.

Figure 6: Results of Sensitivity Analysis



Across the eight combinations of model type and population size, we see very minor differences in the fitted model across the original model priors, less informative priors, and more informative priors. This implies that the prior specifications had a minimal effect on the resulting estimates, an effect that is further minimized when the population size is larger. This outcome highlights the model's robustness across various sets of priors and suggests that the choice of priors have minimal influence on the overall results in the presence of a complete dataset of migration rates over age.

6. Illustrations and Examples

In this section, we work through two examples where we fit our Bayesian Rogers-Castro models on migration data obtained from the 2019 American Community Survey (ACS) through IPUMS USA (IPUMS 2019). The 2019 ACS data provides responses from the American population using a complex sampling design that results in a sampling fraction of around 0.6% (United States Census Bureau 2021). For each respondent, we have the age, state of residence, and the state of residence 1 year prior to the survey. This provides enough information to fit the model for domestic out-migration using either the Poisson or Normal model for each state. We aggregate the microdata to obtain this information while accounting for the weights from the complex sampling design.

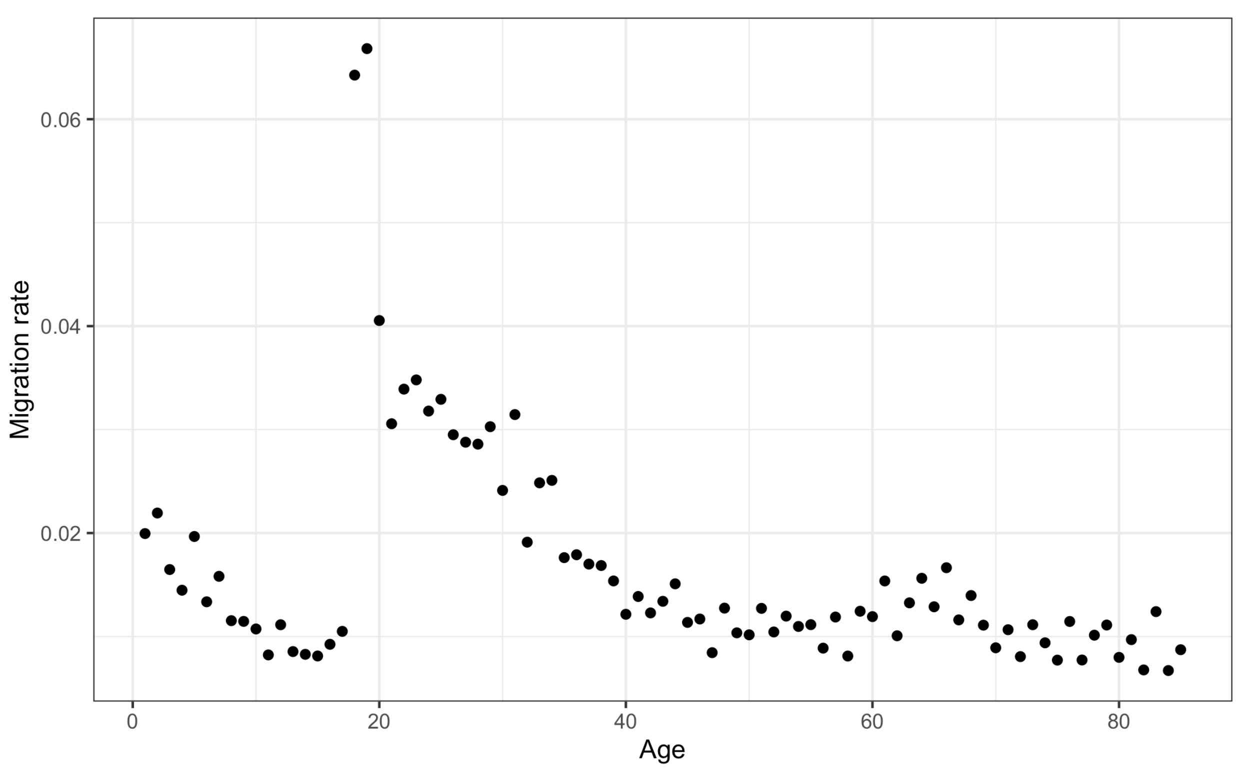
Example 1: Domestic Out-Migration from California, Comparing Poisson and Normal Models

Figure 7a shows the age-specific migration rates for domestic out-migration from the state of California in 2019 for all ages from 1 to 85. From this visualization, it appears appropriate to fit the 11-parameter model that includes the preworking age, working age, and retirement components only. Figures 7b and 7c show the model results from the Poisson model and Normal model respectively, and Table 2 shows the parameter estimates from both models.

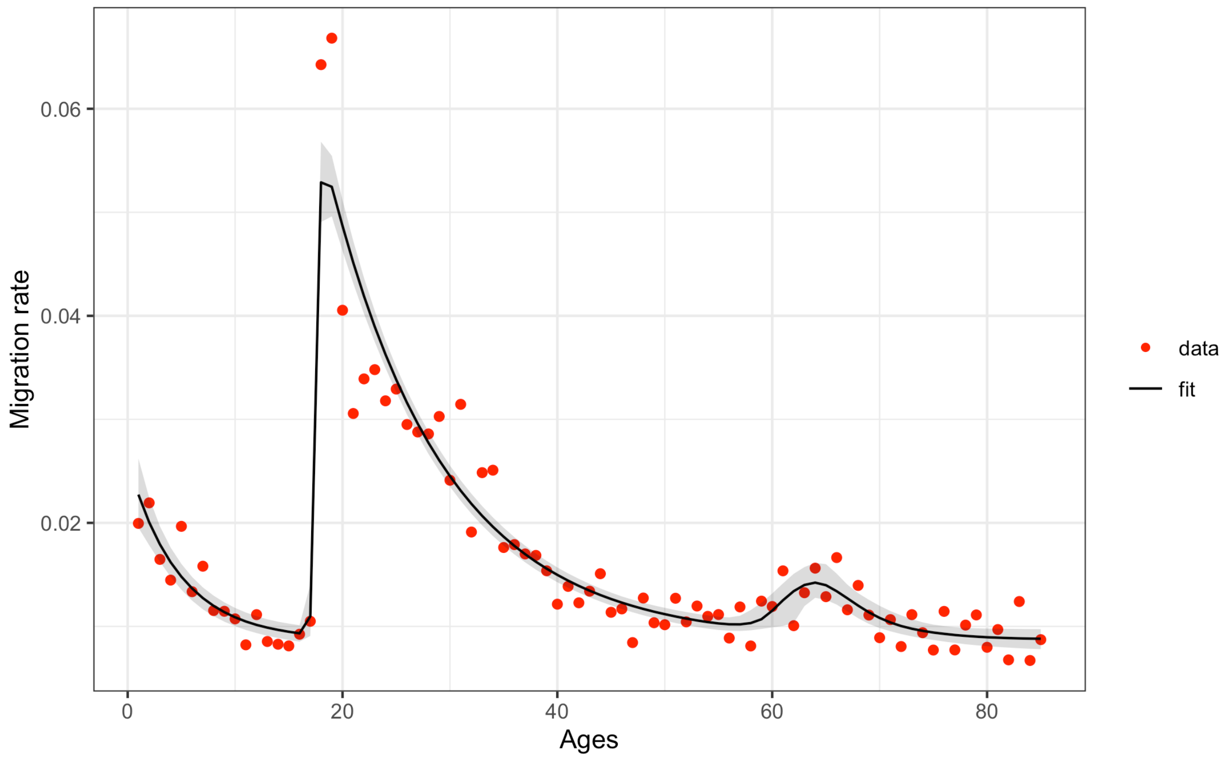
When fitting the Normal model, we specified the standard deviation in the likelihood (σ) to be the standard deviation implied by the Poisson model. In the Poisson specification, the observed number of migrants is assumed to be Poisson distributed with a rate equal to the underlying migration rate multiplied by the size of the population at risk: $y_i \sim Poisson(m_i \times O_i)$, where y_i is the observed number of migrants, m_i is the true age-specific rate, and O_i is the age-specific sample size. The normal approximation to the Poisson is $y_i \sim N(m_i \times O_i, m_i \times O_i)$ or equivalently $r_i \sim N(m_i, m_i / O_i)$ since r_i is the observed migration rate and $r_i = y_i / O_i$. As such, a good choice for σ would be the mean of $\sqrt{r_i/O_i}$ across all age groups i, which is how we calculated the σ in this example. Of course, this technique requires information on the age-specific population or sample sizes which may not be convenient if one only has data on observed age-specific migration rates. If age-specific population sizes or sample sizes are not available, this formula for σ won't be useful as a direct formula, but may still be helpful in determining appropriate values for σ.

Figure 7: Out-migration from California using American Community Survey, 2019

(a) Raw data for out-migration from California, 2019

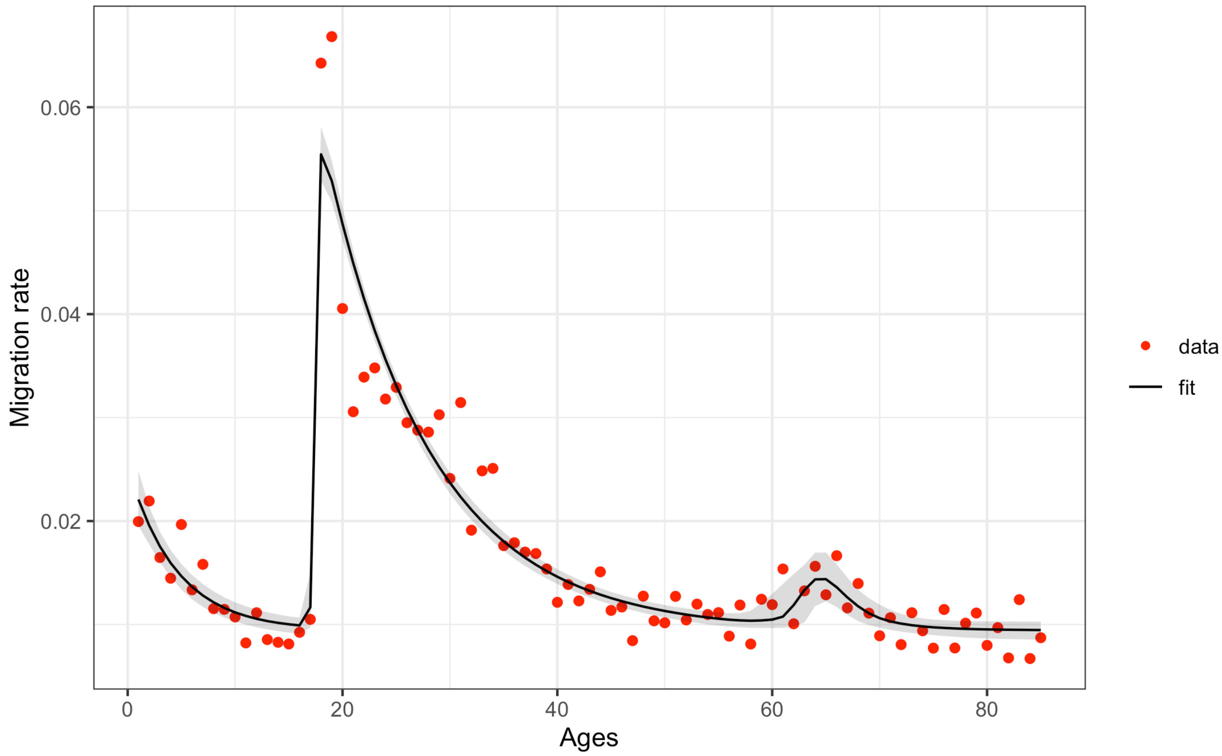


(b) Model result from Poisson model for out-migration from California, 2019



Note: Grey bounds show the 95% credible interval.

(c) Model result from Normal model for out-migration from California, 2019



Note: Grey bounds show the 95% credible interval.

Table 2: Out-migration from California using American Community Survey, 2019

Parameter	**Poisson Model** ***Mean (SD)***	**Normal Model** ***Mean (SD)***
α_1 (alpha1)	0.209 (8.8e-04)	0.222 (8.2e-04)
α_2 (alpha2)	0.092 (1.4e-04)	0.101 (1.3e-04)
α_3 (alpha3)	0.357 (3.3e-03)	0.514 (5.5e-03)
a_1	0.017 (4.1e-05)	0.016 (3.9e-05)
a_2	0.051 (3.4e-05)	0.051 (2.5e-05)
a_3	0.013 (4.7e-05)	0.012 (8.3e-05)

μ_2	17.339 (1.1e-03)	17.283 (9.2e-04)
μ_3	64.815 (1.7e-02)	64.818 (2.2e-02)
λ_2	3.791 (7.4e-03)	4.447 (7.4e-03)
λ_3	0.297 (4.2e-03)	0.518 (8.2e-03)
c	0.009 (1.5e-05)	0.009 (1.1e-05)

Example 2: Sensitivity Analysis of Poisson Model for Domestic Out-Migration from California

For illustrative purposes, example 1 showcased both the Poisson and Normal models. In example 2, we further expand and refine the Poisson model from example 1 through a sensitivity analysis. Initially, the model employed default priors from the *mig_estimate_rc* function in the *rcbayes* package, with the corresponding results already presented. To explore alternative scenarios, competing priors were selected. We fitted the model using two sets of competing priors, as outlined in Table 3: one set comprising less informative priors and another set comprising more informative priors. For the less informative priors, we increased the variability in prior distributions significantly, indicating the limited information available on potential parameter values. Additionally, no prior information was provided regarding the ages at which the working age and retirement peaks (the μ parameters) tend to occur. Conversely, the more informative priors were based on peaks identified in exploratory data analysis (Figure 7a). As well, prior knowledge obtained from other fitted Rogers-Castro migration schedules informed the means of the priors.

Table 3: Priors used in sensitivity analysis for out-migration from California, 2019

Original priors	Less informative priors	More informative priors
$\alpha_1 \sim N(0,1)$ $a_1 \sim N(0,0.1)$ $\alpha_2 \sim N(0,1)$ $a_2 \sim N(0,0.1)$ $\mu_2 \sim N(25,1)$ $\lambda_2 \sim N(0,1)$ $\alpha_3 \sim N(0,1)$ $a_3 \sim N(0,0.1)$ $\mu_3 \sim N(65,1)$ $\lambda_3 \sim N(0,1)$ $c \sim N\left(\min_i \{r_i\}, 0.1\right)$	$\alpha_1 \sim N(0,10)$ $a_1 \sim N(0,10)$ $\alpha_2 \sim N(0,10)$ $a_2 \sim N(0,10)$ $\mu_2 \sim N(1,100)$ $\lambda_2 \sim N(0,10)$ $\alpha_3 \sim N(0,10)$ $a_3 \sim N(0,10)$ $\mu_3 \sim N(1,100)$ $\lambda_3 \sim N(0,10)$ $c \sim N(0,10)$	$\alpha_1 \sim N(0.1,0.5)$ $a_1 \sim N(0.1,0.5)$ $\alpha_2 \sim N(0.1,0.5)$ $a_2 \sim N(0.1,0.1)$ $\mu_2 \sim N(19,0.5)$ $\lambda_2 \sim N(0.1,0.5)$ $\alpha_3 \sim N(0.1,0.5)$ $a_3 \sim N(0.1,0.1)$ $\mu_3 \sim N(66,0.5)$ $\lambda_3 \sim N(0.1,0.5)$ $c \sim N\left(\min_i \{r_i\}, 0.1\right)$

Figure 8: Comparison of Model Results based on different priors

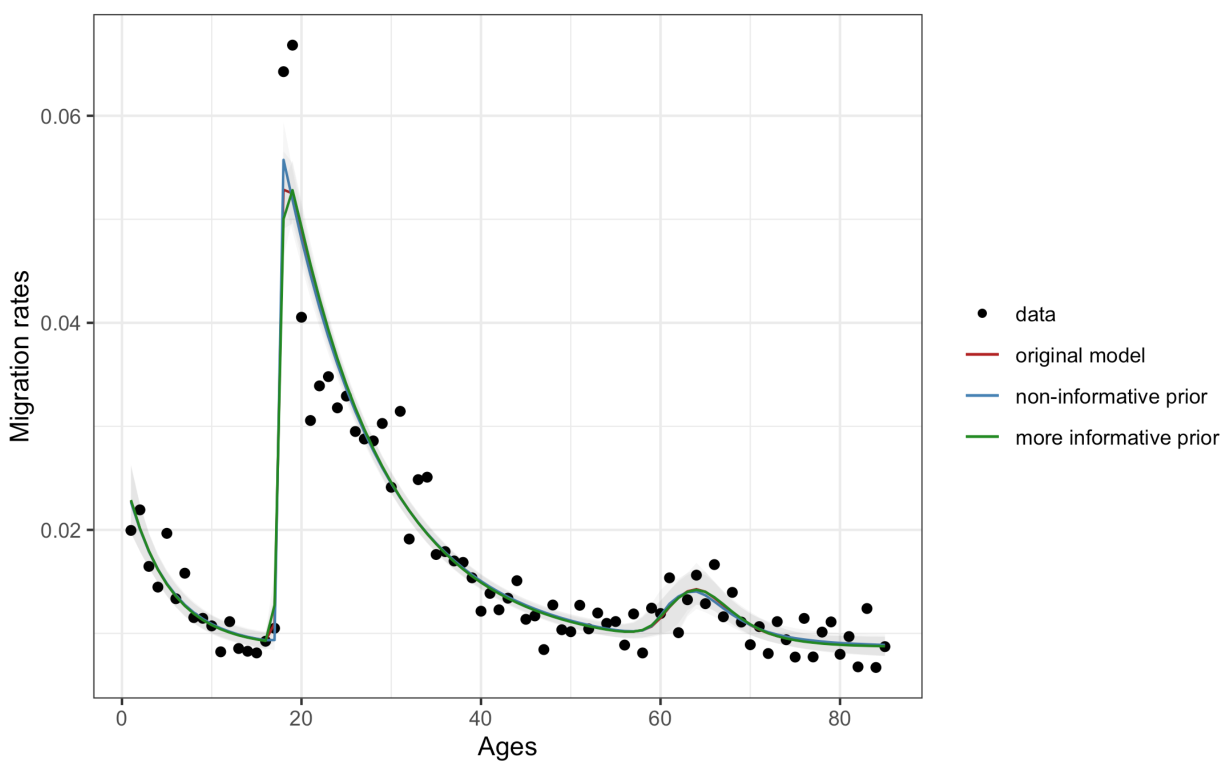


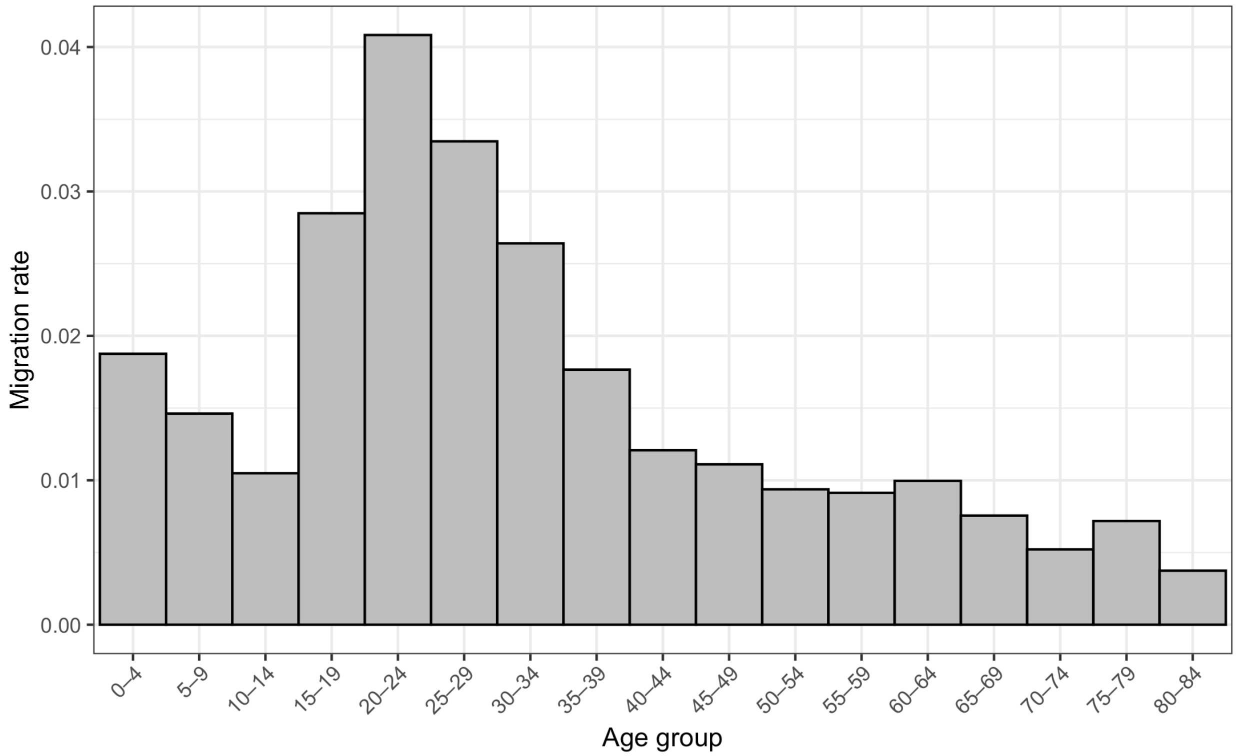
Figure 8 presents the results obtained from the original model as well as the competing models. The plot reveals striking similarity between the estimates derived from the original and competing models. This observation suggests that the prior specifications exerted minimal influence on the final estimates. Despite the utilization of less informative and more informative priors in the sensitivity analysis, the resulting estimates remain nearly indistinguishable.

Example 3: Domestic Out-Migration in Texas, Using data based on 5-year age groups

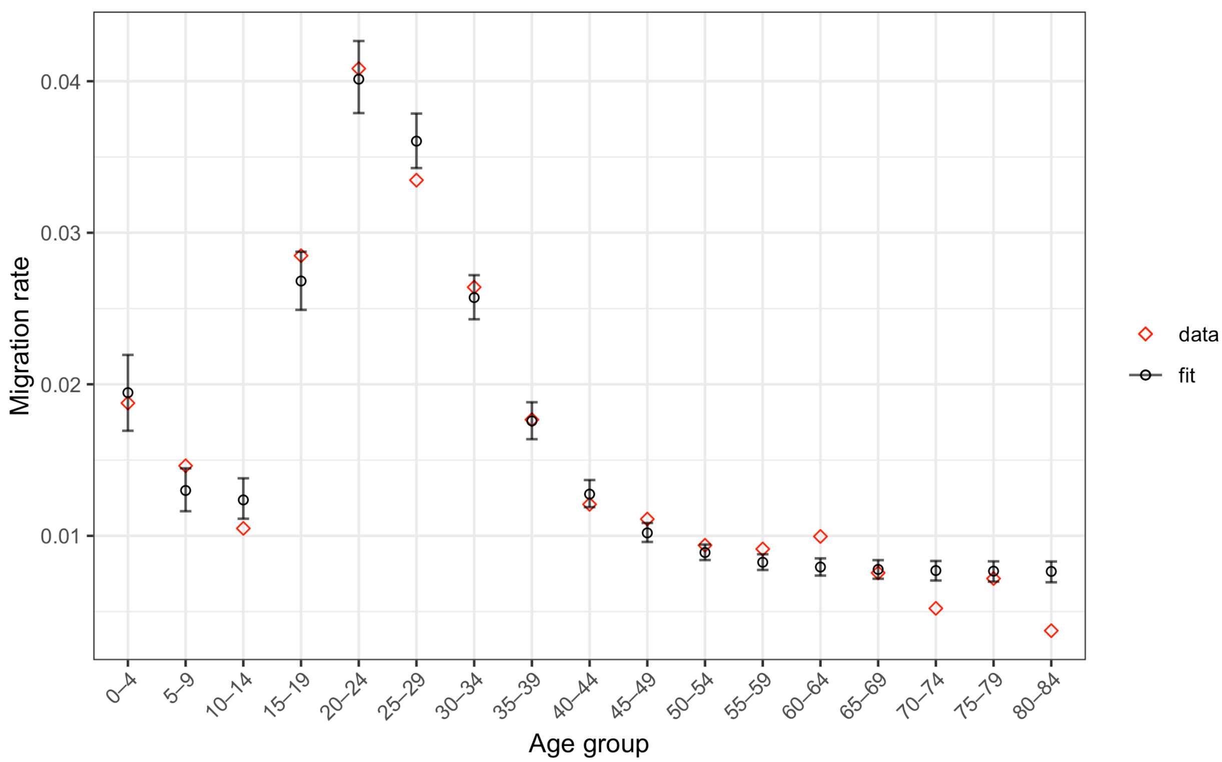
In this example, we demonstrate that we can use this model even when age-specific data comes in the form of 5-year age groups. To do this, we collapse Texas in-migration data into the groups 0–4, 5–9, …, 80–84. Figure 9a shows the migration rates for each age group, which suggests that we should fit the 7-parameter model with preworking age and working-age components. Again, we fit the Poisson model by inputting a vector of 5-year age group midpoints, a vector of migrants for each 5-year age group, and a vector of sample sizes for each 5-year age group. Figure 9b shows the model results.

Figure 9: Out-migration from Texas using American Community Survey, 2019

(a) Raw data for out-migration to Texas, 2019



(b) Model result from Poisson model for out-migration from Texas, 2019



Note: Error bars show the 95% credible interval.

7. Conclusions

This paper proposes two models for fitting the Rogers-Castro migration schedule within a Bayesian framework. We propose one model that accepts migration data as age-specific counts and population sizes, and another that accepts migration data as age-specific rates. Functions for fitting these models are available through the R package *rcbayes*, with the goal of improving accessibility of this methodology to demographers.

The benefits of using these models to fit Rogers-Castro migration schedules are fourfold. First, this method can help identify and avoid some of the issues with existing approaches. Challenges of existing approaches include using optimizers that may stop at local optima, requiring good guesses as initial conditions in order to achieve convergence, and algorithms being very sensitive to those initial conditions. Throughout this paper and R package, we provide tools to conduct a sensitivity analysis to determine the model's robustness to selected prior specifications. Additionally, although convergence is not a guaranteed with our model, we find that convergence issues with the models are almost exclusively with the Poisson model. Even then, using this Bayesian model with additional package utilities (such as *init_rc*) helps alleviate this issue. Second, using Bayesian approaches to fit demographic models such as this are particularly useful in contexts of data-sparsity. Third, our proposed technique is implemented in R which is free, extensible, and widely used in contemporary demographic research. Fourth, the model implementation is provided through the *rcbayes* package and is available on CRAN and maintained in an open GitHub repository, which further promotes accessibility and reproducibility. Further avenues for work on these models can include the inclusion of augmented versions of the Rogers-Castro migration schedule (Wilson 2010) or a hierarchical version of the models which would lend well to fitting schedules for several geographic regions or population subgroups.

References

Bates, J., and Bracken, I. (1982). Estimation of migration profiles in England and Wales. *Environment and Planning A: Economy and Shape*: 14(7): 889-900.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76(1): 1–32.

Congdon, P. (2008). Models for migration age schedules: a Bayesian perspective with an application to flows between Scotland and England. In: Raymer, J., and Willekens, F. (ed.). *International migration in Europe: Data, models and estimates*. West Sussex: John Wiley & Sons, Ltd: 175-192.

Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: demonstrations using an interactive Shiny App. Frontiers in psychology, 11, 608045.

George, M.V. (1994). Population projections for Canada, provinces and territories, 1993–2016. Ottawa: Statistics Canada, Demography Division, Population Projections Section.

Hofmeyer, B.E. (1988). Application of a mathematical model to South African migration data, 1975-1980. *Southern African Journal of Demography* 2(1): 24-28.

Kawabe, H., Hirosima, K., Inoue, T., Hayase, Y., Sato, K., Aihara, Y., and Ishizaki, E. (1990). Migration rates by age group and migration patterns: Application of Roger's migration schedule model to Japan, the Republic of Korea and Thailand. Tokyo: Statistical Research Department, Institute of Developing Economies.

Little, J., and Dorrington, R. (2013). The multi-exponential model migration schedule. In: Moultrie, T., Dorrington, R., Hill, A., Hill, K., Timæus, I., and Zaba, B. (ed.). *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population: 390-402.

Preston, Samuel, Patrick Heuveline, and Michel Guillot. (2000). *Demography: Measuring and Modeling Population Processes*. UK: Wiley-Blackwell.

Raymer, J., and Rogers, A. (2007). Using age and spatial flow structures in the indirect estimation of migration streams. *Demography* 44(2): 199-223.

Raymer, J., and Rogers, A. (2008). Applying model migration schedules to represent age-specific migration flows. In: Raymer, J., and Willekens, F. (ed.). *International migration in Europe: Data, models and estimates*. West Sussex: John Wiley & Sons, Ltd: 175-192.

Rogers, A. and Castro, L.J. (1981). Model Migration Schedules. Research Report RR81-30. Laxenburg: International Institute for Applied Systems Analysis.

Rogers, A., Castro, L.J., and Lea, M. (2005). Model migration schedules: Three alternative linear parameter estimation methods. *Mathematical Population Studies* 12(1) 17-38.

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

Rogers, A., Little, J., and Raymer, J. (2010). *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. New York: Springer Science & Business Media.

Rogers, A., Racquillet, R., and Castro, L.J. (1978). Model migration schedules and their applications. *Environment and Planning A: Economy and Shape* 10(5): 475-502.

Rogers, A., and Raymer, J. (1999). Fitting observed demographic rates with the multiexponential model schedule: An assessment of two estimation programs. *Review of Urban & Regional Development Studies* 11(1): 1-10.

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., Sobek, M., (2021). IPUMS USA: Version 11.0 [dataset]. Minneapolis: IPUMS.

Ruiz-Santacruz, J.S. (2021) MigraR package: Performing Rogers and Castro's multi-exponential models for migration estimation. *Romanian Statistical Review*.

Warnes, A.M. (1992). Age-related variation and temporal change in elderly migration. In: Rogers, A. (ed.). *Elderly migration and population redistribution*. London: Belhaven Press: 35-55.

United Nations. (2019). World Population Prospects 2019: Methodology of the United Nations Population Estimates and Projections.

United States Census Bureau (2021). American Community Survey [electronic resource]. USA: United States Census Bureau. https://www.census.gov/acs.

Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic research* 23: 191-222.

Yeung, J., Alexander, M., and Riffe, T., (2021). rcbayes: Estimate Rogers-Castro Migration Age Schedules with Bayesian Models. R package version 0.1.0.