The Annals of Regional Science (2021) 67:359–383
https://doi.org/10.1007/s00168-021-01051-4

ORIGINAL PAPER

# Smoothing destination-specific migration flows

**Sigurd Dyrting**[1] · **Andrew Taylor**[1]

Received: 17 September 2020 / Accepted: 23 January 2021 / Published online: 12 February 2021


## Abstract

Accurately estimating age profiles for destination-specific migration is requisite to understanding the determinants of population growth and projecting future change as migration is the primary growth determinant for most regions. In Australia, place-to-place flows based on the age profile of migration derived from census data are commonly used to empirically estimate destination-specific internal migration. However, such flows are heterogeneous and census data is imperfect for accurately generating migration-age profiles. Demographers have addressed this by developing a range of methods for smoothing migration probabilities. These address smoothing on a bi-regional basis, primarily with one destination–origin pairing. We propose a non-parametric method for smoothing destination-specific migration probabilities which can be applied to multi-regional smoothing and is within the generation–distribution framework of Rogers et al. (Environ Plan A 34:341–359, 2002). We demonstrate that, if total age-specific out-migration has already been estimated, smoothing destination-specific migration ratios provides a solution to imperfect input data. Using the example of Australian interstate migration, we show how the method can give an accurate fit to the migration ratio profile across high-curvature ages and a good treatment of sample noise both when the population at risk is low, such as at advanced ages, and when the destination has a low conditional probability of migration. An implementation of the method is available as an Excel add-in.

**JEL Classification** J6 · C1 · C8

✉ Sigurd Dyrting
sigurd.dyrting@cdu.edu.au

Andrew Taylor
andrew.taylor@cdu.edu.au

[1] Northern Institute, Charles Darwin University, Darwin, NT, Australia

## 1 Introduction

In-migration to states or territories, cities, and towns in developed nations is, on the whole, the primary determinant for population growth and change (Bell et al. 2015). Accurately estimating in-migration probabilities for specific destinations is therefore important to understanding subnational demographic processes (Smith et al. 2013). Flows of migrants and changes in these between origin and destination communities are, in many cases, more important than changes induced by fertility behaviours (Willekens 2016). Consequently, accurately plotting age profiles of migration to destinations is important to understanding demographic changes and in projecting future migration flows between specific origins and destinations (Rogers 1975, 1986). It is also the basis for comparative work on migration across regions and nations, as well as projecting how these might change over time (Bernard and Bell 2015).

Place-to-place flows are commonly used as the empirical basis for estimating destination-specific internal migration in Australia. Source data are invariably from the 5-yearly Census of Population and Housing from which the demographer can, in theory, pair origins and destinations at fine-grained geographic levels to provide age-specific migration profiles (ABS 2016). Nevertheless, in spite of the commonalities exhibited in the shape of migration profiles across countries and internally [for example, as demonstrated by Rogers and Castro (1981)], the heterogeneity of age-specific migration flows to specific destinations and the imperfectness of census data in accurately depicting underlying profiles remain as ongoing concerns for demographers (Baffour and Raymer 2019).

For census data, one of the most significant issues is the temporal point-in-time capture of movements which negates moves that occurred between the temporal capture points. Specifically, for Australia two questions are included in the census to determine where the respondent lived five years ago and one year ago. These are compared to the current place of usual residence, also provided by respondents, and coded to determine whether a migration event has occurred subsequent to one or five years ago. Standardised geographic frameworks are applied to isolate the spatial specificities of the migration event and these feed into population estimates by age and gender (ABS 2019) as well as projection processes.

While important to demographic analysis, the point-in-time capture of migration is limited (to one- and 5-year intervals) and masks underlying dynamicism in the flows and age profiles for migration to and from specific destinations. Not least, the absolute number of migration events is under-captured by an unknown amount since an individual may undertake multiple moves within the one- or 5-year interval, aside from the ones captured at those points in time (Courgeau 1973). This is pertinent for understanding migration in Australian because, during their life course, Australians undertake more moves on average than residents of other developed nations[1] (Bell et al. 2015, 2017).

[1] In Bell et al. (2015)'s league table of forty-five countries ranked by 1-year aggregate crude migration intensity Australia is the seventh highest, below Iceland, Finland, Zambia, Kenya, Denmark, and the USA.

The one- and 5-year intervals for census migration data also creates problems for demographers seeking to compare or convert across time or areas. For example, absolute numbers of migration flows captured by the 1-year variable and multiplied by five are generally significantly larger than captured in the 5-year variable. This difference is due to multiple moves made by individuals during the longer period (Rees 1977). Demographers call this the "1-year/5-year" problem, reflective of variations between migration probabilities collected over one- and 5-year intervals and hindering direct comparisons between the two. To address this issue, some demographers favour migration ratios, the probability of migrating to a particular destination conditional on out-migrating from a given origin, for their relative stability in derivation from point-in-time intervals (Rogers et al. 2003).

Studies have demonstrated further weaknesses exist in census migration data including in the accuracy of destination migration which is captured for infants born within the transition interval (Rees et al. 2000) and for older migrants (Wilson and Terblanche 2018; Wilson 2020). In addition, age and gender profiles in census migration data for more rural and remote destinations are known to be more volatile, partly as a result of the issues above and combined with temporal unpredictability due to their relatively small populations (Peters et al. 2016). As a key input to modelling, the cumulative impacts of deviations from the underlying age profiles for migration are to reduce the accuracy of comparative migration metrics and their inputs to other forms of population modelling like projections.

In theory, some of the shortcomings in census data might be overcome through application of administrative datasets capturing migration. However, in Australia at least, while the landscape of unit record administrative data provision for such purposes is improving, there are currently no nationally consistent datasets (with sufficient coverage and longitudinal scope) for application to this purpose. Consequently, demographers continue to develop methods for generating accurate destination-specific profiles from noisy and incomplete data (Rogers et al. 2010).

A growing number of studies for modelling and smoothing migration profiles are evident (for example, Rogers et al. (1978), Rogers and Castro (1981), Rogers and Watkins (1987), Congdon (2008), Wilson (2010), Bernard and Bell (2015), Wilson (2020), Dyrting (2020)). The basis for attempts to estimate smooth origin–destination-age-specific migration intensities are counts of movers by destination from an origin population, both of which are available by single year of age from census data. Modelling migration involves a multinomial process within which smooth intensities can be estimated using a number of methods. Willekens (2008), for example, utilised the maximum likelihood method which Hachen (1988) highlighted can be approached from either a competing-risk perspective, where out-migration is estimated separately for each destination, or from a generation–distribution perspective (Rogers et al. 2002), where out-migration and migration ratios are estimated separately. Both approaches have been used to prepare inputs to models for projecting subnational populations (Campbell 1996; Nash 2020).

P-splines are a powerful tool for smoothing and have been applied to estimating mortality (Currie et al. 2004; Camarda 2012; Gonzaga and Schmertmann 2016), and the generation component of out-migration (Dyrting 2020). Our aim here is to extend the method to more accurately smooth the distribution component of

migration than methods currently employed. Our approach is motivated by Rogers et al. (2002)'s observation that, while migration probabilities are strongly age-dependent, migration ratios are less so. On this basis, the problem of separating signal from noise, which all smoothing methods seek to solve, should be more effective for ratios than for probabilities.

In the next section, we give a review of transition-style migration probabilities and ratios and summarise the problem of removing irregularities in the age profile through smoothing. In Sects. 3 and 4 we introduce the multinomial model and the P-spline method applied to ratios, demonstrating its solution by iterated linear regressions. In Sect. 5, we apply the method to smoothing the distribution component of interstate migration for Australia. In Sect. 6, we illustrate how it can be combined with a method for smoothing the generation component of out-migration to smooth destination-specific migration probabilities. In Sect. 7 we show how the method can be used to directly compare ratios for 1-year and 5-year migration intervals.

## 2 Migration ratios

Transition type migration data, such as is collected by the Australian Census for Population and Housing, consists of observations

$$ {}_nM^j = \begin{bmatrix} {}_nM_0^j \\ \vdots \\ {}_nM_\omega^j \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} N_0 \\ \vdots \\ N_\omega \end{bmatrix} \tag{1} $$

of ${}_nM_x^j$ movers of age $x+n$ to destination $j$ from an initial population $N_x$ of age $x$ of a specific origin. Raw out-migration probability ${}_n\tilde{m}$ and migration ratios ${}_n\tilde{c}^j$ are given by the fractions

$$ {}_n\tilde{m} = \frac{{}_nM}{N} \quad \text{and} \quad {}_n\tilde{c}^j = \frac{{}_nM^j}{{}_nM}, \tag{2} $$

where here and in the following all matrix operations and functions act elementwise and

$$ {}_nM = \sum_{j=1}^{d} {}_nM^j \tag{3} $$

is the vector of total movers, and $d$ is the number of possible destinations.

Within the generation–distribution framework ${}_n\tilde{m}$ is the observed probability of out-migrating from the origin area, and ${}_n\tilde{c}^j$ is the observed probability of migrating to destination $j$ conditional on out-migrating. We see from Eq. (2) that $N$ is the population exposed to the risk of out-migrating and ${}_nM$ is the population of out-migrants exposed to the risk of in-migrating to $j$. Since both of these populations are finite, the observed probabilities will have elements of sample noise which we seek

to remove using smoothing. In the following we assume that a good estimate of total out-migration ${}_nm$ has been obtained and address the problem of finding smooth vectors ${}_nc^j$ that fit the observed ratios ${}_n\tilde{c}^j$.

## 3 A multinomial model of migration

Assuming a multinomial model of migration counts (Willekens 2008) one can show that the log likelihood of observing ratios ${}_n\tilde{c}^j$ when the underlying ratios are ${}_nc^j$ is

$$\mathscr{L}_c = {}_nM' \cdot \sum_{j=1}^{d} {}_n\tilde{c}^j \log({}_nc^j), \tag{4}$$

where $A \cdot B$ denotes matrix multiplication and $A'$ is the transpose of $A$.

It is difficult to maximise the above form of the log likelihood function because there is an auxiliary condition that the migration ratios must sum to unity. To handle this condition, it is useful to express ${}_nc^j$ in terms of conditional ratios ${}_na^j$ defined by

$${}_nc^j = \begin{cases} {}_na^1, & j = 1 \\ {}_ns^j \times {}_na^j, & j = 2, \ldots, d-1 \\ {}_ns^d, & j = d \end{cases} \tag{5}$$

where ${}_ns^j$ is the product

$${}_ns^j = \prod_{1 \le k < j} (1 - {}_na^k). \tag{6}$$

Observed conditional ratios ${}_n\tilde{a}^j$ are defined similarly. ${}_na^j$ is the probability of migrating to $j$ conditional on not migrating to destinations $1, \ldots, j-1$. Let ${}_nK^j$ be the number of movers to destinations with index $j$ or greater

$${}_nK^j = \sum_{k=j}^{d} {}_nM^k, \tag{7}$$

then the log likelihood function can be rewritten as

$$\mathscr{L}_c = \sum_{j=1}^{d-1} \mathscr{L}_j, \tag{8}$$

where

$$\mathscr{L}_j = ({}_nK^j)' \cdot y_j, \tag{9}$$

and

$$y_j = {}_n\tilde{a}^j \log {}_n a^j + (1 - {}_n\tilde{a}^j) \log(1 - {}_n a^j). \tag{10}$$

The derivation of Eq. (8) is given in Sect. A.2. Written in this form the likelihood is easier to maximise as we only need to impose the condition $0 \leq {}_n a^j \leq 1$ on the conditional ratios. Finally, in order to give ratios a common representation which is independent of the interval $n$ we express the $n$-year conditional ratios ${}_n a^j$ in terms of implied ratios at 1-year intervals $a^j$

$${}_n a^j = {}_n T^j \cdot a^j, \tag{11}$$

where the matrix ${}_n T^j$ is given iteratively by

$$\begin{aligned} {}_n T^1 &= {}_n U, \\ {}_n T^j &= \operatorname{diag}\Big(\tfrac{1}{1-{}_n a^{j-1}}\Big) \cdot {}_n T^{j-1} \cdot \operatorname{diag}\big(1 - a^{j-1}\big). \end{aligned} \tag{12}$$

Here ${}_n U$ is the matrix with elements

$${}_n U_{xr} = \begin{cases} 0 & r < x \\ \big(\prod_{x \leq k < r} (1 - m_k)\big) m_r / {}_n m_x & x \leq r < x + n \\ 0 & r \geq x + n \end{cases}, \tag{13}$$

and $m_k$ and ${}_n m_x$ are probabilities obtained by smoothing total out-migration (Dyrting 2020). The derivation of Eq. (11) is given in Sect. A.3. Our strategy is now to smooth conditional ratios by maximising $\mathscr{L}_j$ sequentially as a function of $a^j$ only.

## 4 Penalised splines

In this section we use penalised splines (P-splines) to smooth conditional ratios (Eilers and Marx 1996). Since we will be smoothing them sequentially, we drop the destination index $j$ to lighten the notation. Represent implied conditional ratios in terms of B-splines using the functional form

$$\operatorname{logit}(a) = B \cdot \theta. \tag{14}$$

Here $B$ is a matrix of B-spline functions arranged columnwise (de Boor 2001). Conditional ratios are smoothed by maximising the penalised log likelihood function

$$\mathscr{L} = {}_n K' \cdot y - \frac{\lambda}{2} \theta' \cdot D_k' \cdot D_k \cdot \theta, \tag{15}$$

where $D_k$ is the $k$-order difference matrix and $\lambda > 0$ is a penalty parameter. In principle this equation could be solved for $\theta$ using a multivariate optimisation routine, but because the number of B-spline nodes is potentially large we need an alternate solution method. Assuming the maximum of the penalised log likelihood occurs at a stationary point we get a system of nonlinear equations for $\theta$ which can be solved

by iterative linear regressions. Let $\bar{\theta}$ be the current approximation to the B-spline weights. The updated value $\theta$ is the solution to

$$Q\left(\bar{\theta}\right) \cdot \theta = b(\bar{\theta}), \tag{16}$$

where

$$Q(\theta) = G' \cdot W(\theta) \cdot G + \lambda D_k' \cdot D_k, \tag{17}$$

$$b(\theta) = G' \cdot V \cdot \left({}_n\tilde{a} - {}_n a\right) + G' \cdot W(\theta) \cdot G \cdot \theta, \tag{18}$$

and

$$W(\theta) = \mathrm{diag}\left({}_n a(1 - {}_n a)\, {}_n K\right). \tag{19}$$

The derivation of this iteration and the expression for $G$ are given in Appendix 2.

Smoothness is controlled by the penalty parameter $\lambda$. The higher the value the smoother the ratio $a$. The penalty can be specified explicitly or chosen automatically by minimising one of a number of measures that seek to balance the decreased fitting error against the increased effective number of parameters as $\lambda$ is made smaller. Two popular measures are the Akaike information criterion (Akaike 1974)

$$\mathrm{AIC}(\lambda) = \mathrm{dev} + 2 \times \mathrm{dim}, \tag{20}$$

and the Bayesian information criterion (Schwarz 1978)

$$\mathrm{BIC}(\lambda) = \mathrm{dev} + \mathrm{dim} \times \log(1 + \omega), \tag{21}$$

where

$$\mathrm{dev}(\theta, \lambda) = -2 \times {}_n K' \cdot y \tag{22}$$

is the deviance and

$$\mathrm{dim}(\theta, \lambda) = \mathrm{tr}(H) \tag{23}$$

is the effective dimension of $\theta$ calculated using the trace of the hat matrix of the linearised problem

$$H = \left(G' \cdot W \cdot G + \lambda D_k' \cdot D_k\right)^{-1} \cdot G' \cdot W \cdot G. \tag{24}$$

From experiments with Australian census data we found that AIC($\lambda$) would often under-smooth and BIC($\lambda$) would often over-smooth. We found that a good compromise was the Akaike information criterion with corrections (Hurvich and Tsai 1989)

$$\mathrm{AICc}(\lambda) = \mathrm{AIC}(\lambda) + 2\frac{\mathrm{dim}(\mathrm{dim} + 1)}{\omega - \mathrm{dim}} \tag{25}$$

## 5 Application to interstate migration: ratios

As an application of the smoothing method outlined above, we consider estimation of the distribution component of Australia interstate migration. This is an important step in the preparation of origin–destination-age-specific migration probabilities which are necessary inputs to a multistate lifetable analysis or population projection model (Rogers 1975). Data from the 2016 Australian Census of Population and Housing were used to calculate raw and smoothed destination-age-specific out-migration ratios for each of Australia's six states and two mainland territories for both 1-year and 5-year intervals. With P-splines the knots should be spaced at intervals small enough such that an unpenalised ratio ($\lambda = 0$) will show more variation than is justified by the data (Eilers and Marx 1996). For most age ranges we found that a knot spacing of approximately three years was sufficiently small. For eastern states, out-migration to the Australian Capital Territory changes rapidly over the age ranges 17 to 19 (1-year ratios) and 12 to 19 (5-year ratios) reflecting its importance as a destination for young adults entering tertiary education. Therefore, over the age interval 12 to 21 we used knots spaced at 1-year intervals and for the remainder of the age range 0 to 90 we used knots at 3-year intervals. Our fits did not change substantially if a finer grid for the knots was used. The results presented are for quadratic B-splines because we found that linear B-splines would occasionally give kinks at the knot points. A linear penalty ($k = 1$) was used, with $\lambda$ determined by minimising AICc($\lambda$). For comparison we have included the results of a linear kernel regression smoothing of the raw migration ratios using a Gaussian kernel (Fan and Gijbels 1996) and the Rule-Of-Thumb bandwidth selector (Ruppert et al. 1995).

Figures 1 and 2 compare raw and smoothed 1-year migration ratios for Australia's largest state, New South Wales, and its least populated mainland territory, the Northern Territory. The complete set of 112 origin–destination ratios for all states and territories is given in Figures S-1 to S-16 of Online Resource 1. Also shown is the 95% confidence interval for observed ratios based on the P-spline fit. As observed by Rogers et al. (2002) ratios do not exhibit as strong a dependence on age as probabilities. Also, apart from a strong constant component there does not appear to be a repeating pattern common across destinations, and yet we do observe a definite variation with age, in particular the presence of a "student peak" for migration to the ACT from New South Wales, Victoria, and Queensland (see Fig. 1, S-2, and S-3). The two smoothing methods give similar results when both the origin population and the destination-specific ratio are large (see the age-specific migration ratio from New South Wales to Queensland). They begin to show differences when dispersion in the raw data increases, especially over advanced ages (see the age-specific migration ratio from the Northern Territory to Tasmania).

Table 1 gives summary statistics for the two smoothing methods: a goodness-of-fit measure given by the deviance

$$\mathrm{dev}_c = 2 \times {}_nM' \cdot \sum_{j=1}^{d} {}_n\tilde{c}^j \log({}_n\tilde{c}^j / {}_nc^j), \tag{26}$$

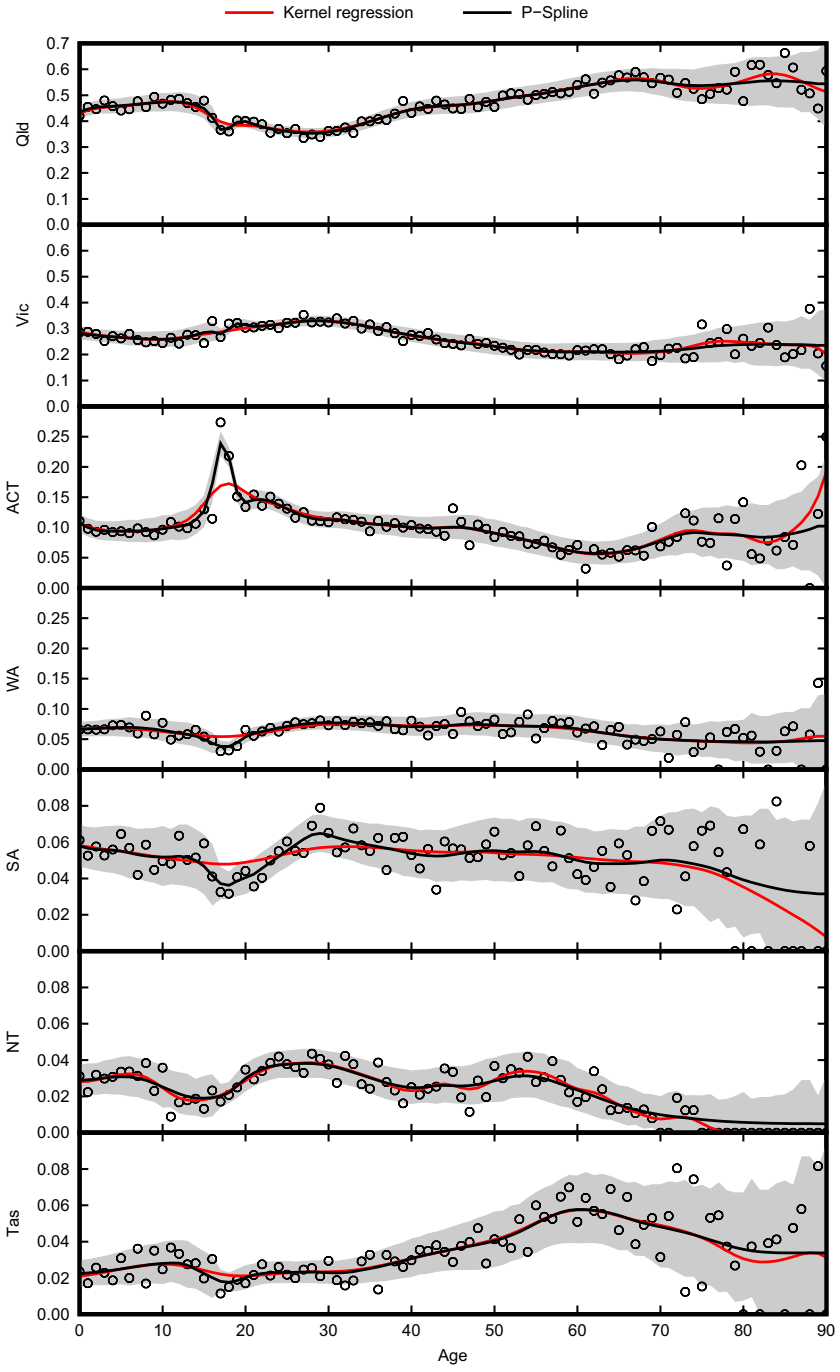


**Fig. 1** New South Wales interstate out-migration ratios 2015–2016 by age in 2015 and destination, two smoothing methods. Grey area, 95% confidence interval for observed ratios based on P-spline fit. *Source*: Based on Australian Bureau of Statistics data

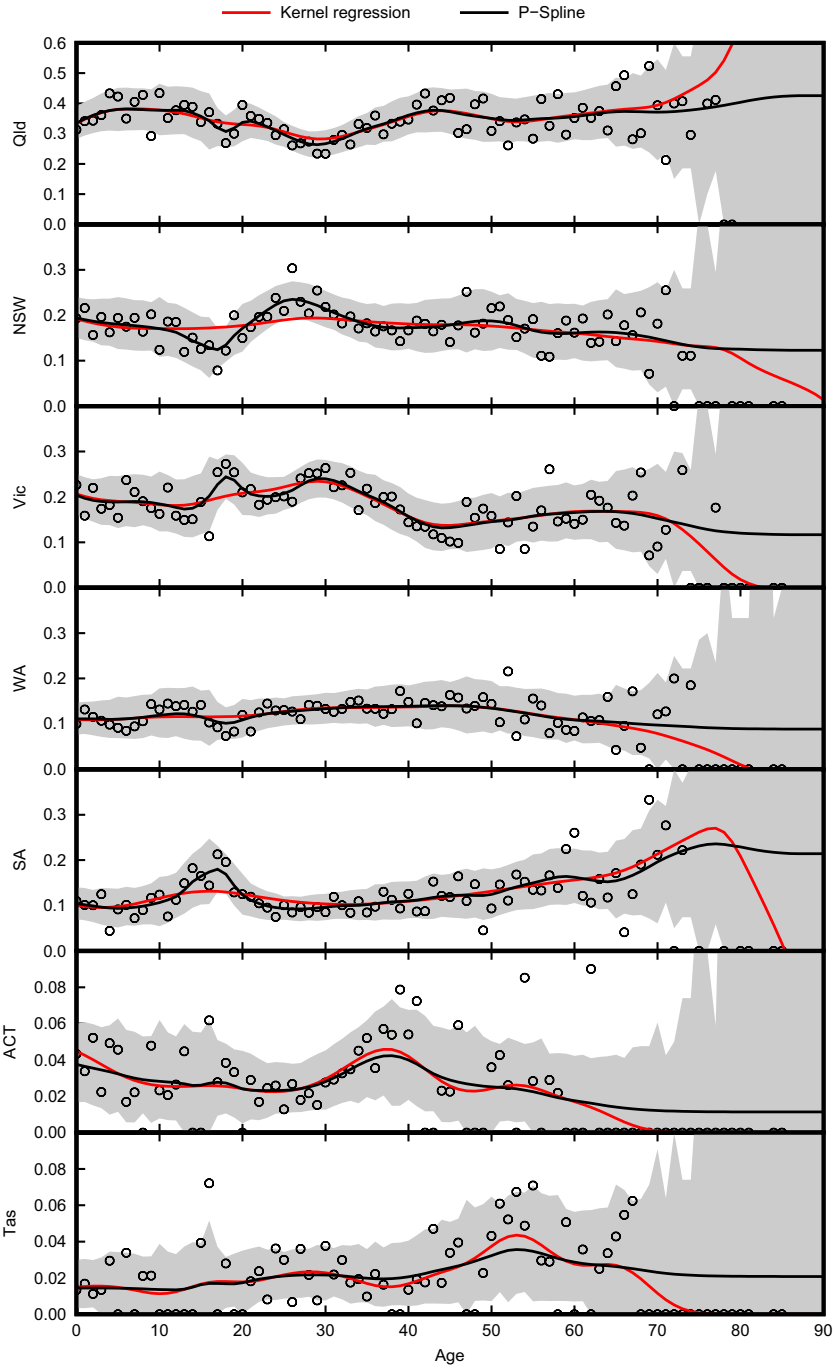


**Fig. 2** Northern Territory interstate out-migration ratios 2015–2016 by age in 2015 and destination, two smoothing methods. Grey area, 95% confidence interval for observed ratios based on P-spline fit. *Source*: Based on Australian Bureau of Statistics data

**Table 1** Summary statistics for two smoothing methods applied to Australian interstate out-migration ratios, 2016

| State | $n$ | $\text{dev}_c$ | | Notes | | Fig. |
|---|---|---|---|---|---|---|
| | | K | P | K | P | |
| NSW | 1 | 1131 | 916 | S | | 1 |
| VIC | 1 | 1086 | 832 | S | | S-2 |
| QLD | 1 | 888 | 878 | S,E | | S-3 |
| WA | 1 | 942 | 877 | E | | S-4 |
| SA | 1 | 1031 | 935 | E | | S-5 |
| TAS | 1 | 1174 | 1061 | | | S-6 |
| ACT | 1 | 1113 | 1161 | E | | S-7 |
| NT | 1 | 922 | 851 | E | | 2 |
| NSW | 5 | 760 | 797 | E | | S-9 |
| VIC | 5 | 988 | 834 | E | | S-10 |
| QLD | 5 | 871 | 820 | S,E | | S-11 |
| WA | 5 | 1221 | 914 | E | | S-12 |
| SA | 5 | 917 | 795 | E | | S-13 |
| TAS | 5 | 884 | 856 | E | | S-14 |
| ACT | 5 | 888 | 996 | E | | S-15 |
| NT | 5 | 750 | 693 | E | | S-16 |

Goodness-of-fit measure $\text{dev}_c$ is the multinomial deviance given by Eq. (26). The Notes column gives the authors' assessment of a fit's deficiencies, if any: *S*, over-smoothing the student peak; *E*, under-smoothing advanced ages. K, kernel regression; P, P-splines. Figures S-2 to S-7 and S-9 to S-16 are given in Online Resource 1

and our assessment of each fit's deficiencies, if any, focussing on two types: Over-smoothing of the student peak and under-smoothing of the profile at ages 80 and over. P-spline has the lowest value of $\text{dev}_c$ for 13 of the 16 state-interval combinations. For the three cases where kernel regression has the lowest deviance its fit shows signs of under-smoothing at senior ages. P-spline performs better because it is able to model the increase in sample variance with decreasing population at risk (the total number of movers) whereas kernel regression assumes it is the same for all ages. Furthermore, kernel regression has over-smoothed the student peak to ACT in the 1-year data for New South Wales, Victoria, and Queensland (Fig. 1, S-2, and S-3 respectively) and the 5-year data for Queensland (Figure S-11).

## 6 Application to interstate migration: probabilities

Once smoothed ratios have been found, all of the destination-specific migration probabilities ${}_nm^j$ are then available to us through the expression

$$ {}_nm^j = {}_nm \times {}_nc^j. \tag{27}$$

An alternate framework for smoothing is the competing-risk approach (Hachen 1988), where age-specific probabilities are smoothed separately for each destination. We compare our generation–distribution approach to two competing-risk smoothing methods: smoothing with kernel regression (Bernard and Bell 2015) and smoothing with Wilson (2010)'s student model migration schedule (student MMS). Data from the 2016 Australian Census of Population and Housing were used to calculate raw and smoothed destination-age-specific out-migration probabilities for each of Australia's six states and two mainland territories for both 1-year and 5-year intervals. The generation component was smoothed using P-TOPALS (Dyrting 2020) and the distribution component taken from Sect. 5. For kernel regression, raw destination-specific migration probabilities were smoothed using linear polynomials, a Gaussian kernel (Fan and Gijbels 1996), and the Rule-Of-Thumb bandwidth selector (Ruppert et al. 1995). For student MMS, destination-specific migration probabilities were smoothed with a three-step process:

Step 1: the model was fitted to total out-migration, and the parameter values saved.
Step 2: for each destination the model was fitted to destination-specific migration probabilities, keeping the profile parameters fixed at their values from Step 1 and only adjusting the level parameters.
Step 3: for each destination all parameters were fitted starting from their Step 2 values.

Figures 3 and 4 compare raw and smoothed 1-year migration probabilities for New South Wales and the Northern Territory. The complete set of 112 origin–destination probabilities for all states and territories are provided in Figures S-17 to S-32 of Online Resource 1. These figures also show the 95% confidence interval for observed probabilities based on the P-spline fit. The origin–destination-specific schedules display a variety of differences in the position and prominence of student, labour, and retirement peaks. Because it is concentrated over a narrow age range, a prominent student peak will not be well fitted by kernel regression, which tends to over-smooth the feature (see New South Wales to ACT in Fig. 3, and Northern Territory to SA in Fig. 4).

Approximating out-migration to a given destination $j$ as a Poisson process it can be shown that the size of the sample noise in ${}_n\tilde{m}^j$ relative to ${}_n m^j$ is $1/\sqrt{N \times_n m^j}$. This implies that the relative size of the sample noise increases as the exposed population $N$ decreases. It also shows that the relative size of sample noise will be larger for destinations with lower probabilities. In each of Figs. 3, 4, and S-17 to S-32 destinations are arranged from top to bottom in order of decreasing probability. We see an increase in the amount of sample noise relative to the level, and when it is large both kernel regression and student MMS can give unrealistic profiles (see Northern Territory to Tasmania in Fig. 4).

Table 2 gives summary statistics for the three smoothing methods: two goodness-of-fit measures dev and $\text{dev}_m$, a shape plausibility measure $\bar{P}$, and our assessment of each fit's deficiencies, if any. The first goodness-of-fit measure is the multinomial deviance

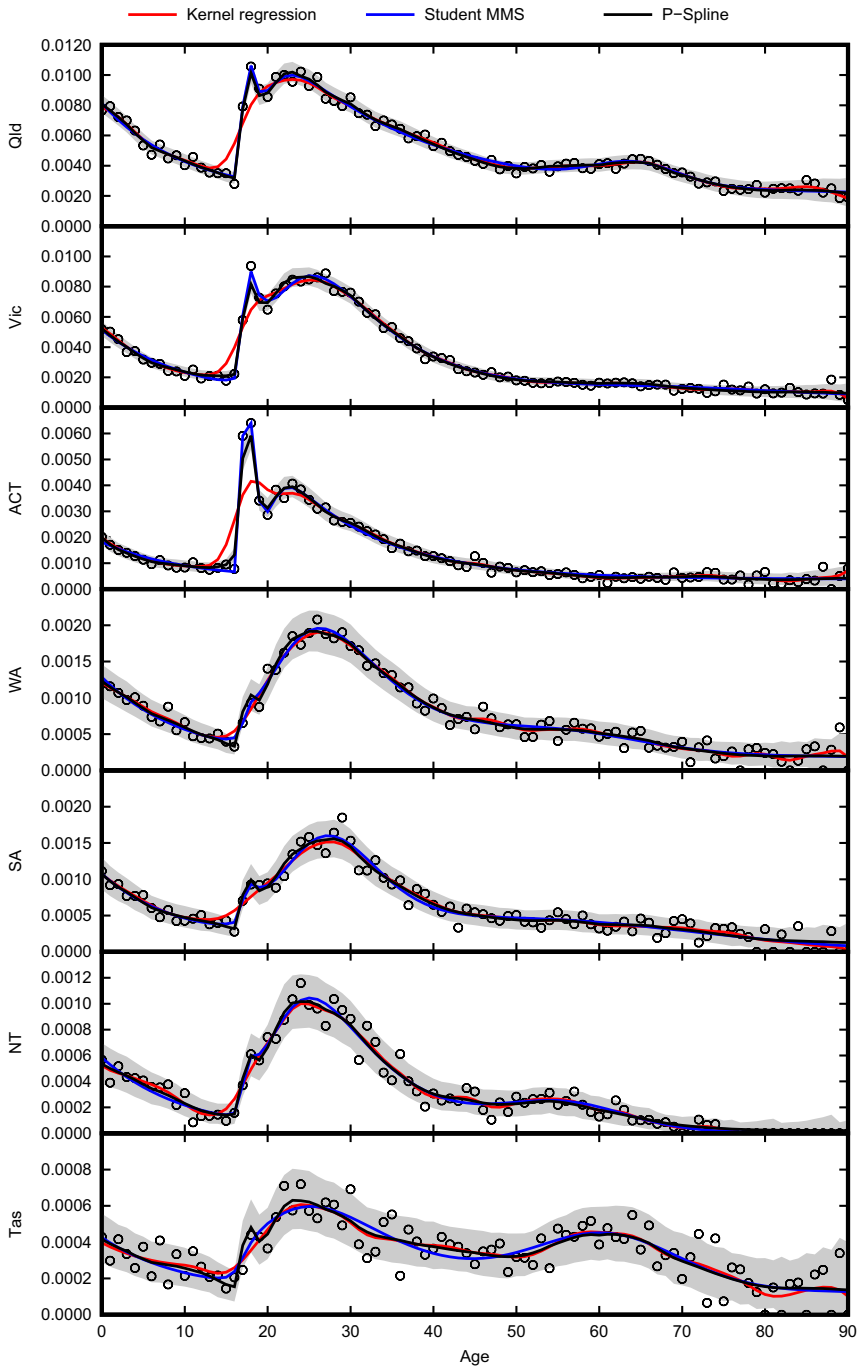


**Fig. 3** New South Wales interstate out-migration probability 2015–2016 by age in 2015 and destination, three smoothing methods. Grey area, 95% confidence interval for observed probabilities based on P-spline fit. *Source*: Based on Australian Bureau of Statistics data

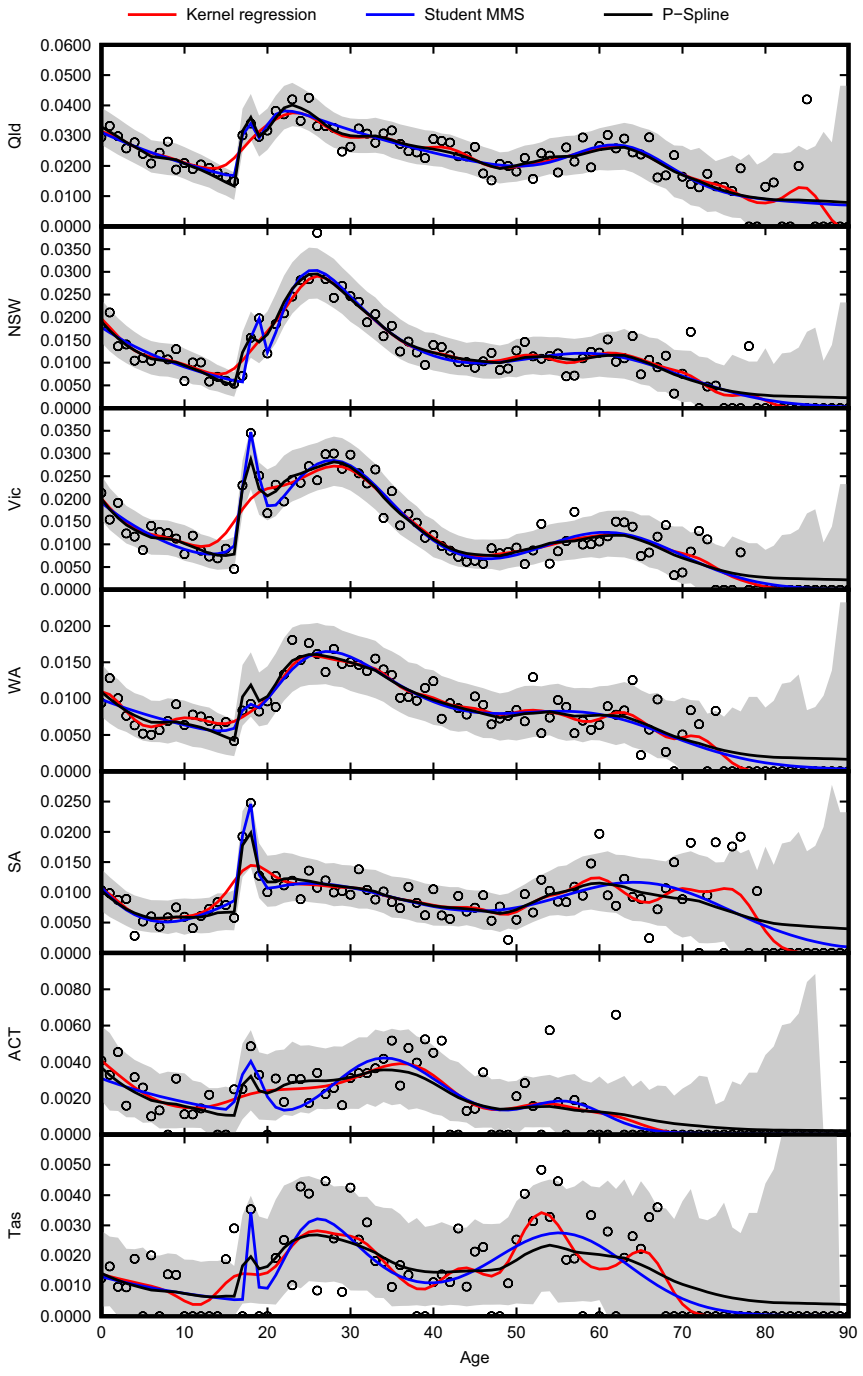


**Fig. 4** Northern Territory interstate out-migration probability 2015–2016 by age in 2015 and destination, three smoothing methods. Grey area, 95% confidence interval for observed probabilities based on P-spline fit. *Source*: Based on Australian Bureau of Statistics data

**Table 2** Summary statistics for three smoothing methods applied to Australian interstate out-migration probabilities, 2016

| State | $n$ | dev | | | $\text{dev}_m$ | | | $\bar{P}$ | | | Notes | | | Fig. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | K | M | P | K | M | P | K | M | P | K | M | P | |
| NSW | 1 | 1611 | 898 | 861 | 856 | 98 | 93 | 17 | 18 | 15 | S | | | 3 |
| VIC | 1 | 927 | 933 | 878 | 255 | 107 | 109 | 12 | 14 | 11 | S | | | S-18 |
| QLD | 1 | 931 | 931 | 835 | 296 | 122 | 98 | 16 | 16 | 14 | S | | | S-19 |
| WA | 1 | 952 | 1026 | 929 | 280 | 161 | 141 | 21 | 20 | 18 | S,E | | | S-20 |
| SA | 1 | 1031 | 1143 | 1,049 | 239 | 164 | 114 | 16 | 15 | 12 | S,E | | | S-21 |
| TAS | 1 | 1166 | 1165 | 1158 | 207 | 130 | 97 | 23 | 25 | 21 | S,X | | | S-22 |
| ACT | 1 | 1146 | 1339 | 1239 | 272 | 166 | 164 | 29 | 17 | 11 | S,X | | | S-23 |
| NT | 1 | 955 | 982 | 998 | 212 | 154 | 147 | 44 | 45 | 29 | S,X | X | | 4 |
| NSW | 5 | 665 | 901 | 649 | 230 | 194 | 86 | 22 | 23 | 22 | | | | S-25 |
| VIC | 5 | 723 | 1039 | 840 | 147 | 207 | 127 | 17 | 18 | 17 | | | | S-26 |
| QLD | 5 | 689 | 1051 | 744 | 177 | 242 | 111 | 17 | 16 | 16 | E | | | S-27 |
| WA | 5 | 686 | 852 | 825 | 119 | 125 | 121 | 25 | 25 | 23 | | | | S-28 |
| SA | 5 | 768 | 1008 | 892 | 132 | 157 | 97 | 16 | 17 | 15 | E | | | S-29 |
| TAS | 5 | 739 | 1146 | 962 | 115 | 155 | 127 | 25 | 21 | 21 | X | | | S-30 |
| ACT | 5 | 770 | 1081 | 1035 | 153 | 198 | 140 | 28 | 29 | 17 | X | | | S-31 |
| NT | 5 | 731 | 913 | 848 | 154 | 171 | 155 | 39 | 32 | 32 | | | | S-32 |

Goodness-of-fit measure dev is the multinomial deviance for all destination-specific migration probabilities given by Eq. (28). Goodness-of-fit measure $\text{dev}_m$ is the binomial deviance for total out-migration probability given by Eq. (29). Quantity $\bar{P}$ is given by Eq. (30) and equals the sum of the percentage that each destination-specific schedule's profile differs from a reference profile. The Notes columns give the authors' assessment of a fit's deficiencies, if any: *S*, over-smoothing the student peak; *E*, under-smoothing advanced ages; *X*, implausible shape. K, kernel regression; M, student MMS; P, P-TOPALS (generation)+P-splines (distribution). Figures S-18 to S-23 and S-25 to S-32 are given in Online Resource 1

$$\text{dev} = 2 \times N' \cdot \left[ (1 - {}_n\tilde{m}) \log \left( \frac{1 - {}_n\tilde{m}}{1 - {}_n m} \right) + \sum_{j=1}^{d} {}_n\tilde{m}^j \log \left( \frac{{}_n\tilde{m}^j}{{}_n m^j} \right) \right] \tag{28}$$

for a joint fit of all destination-specific migration probabilities for a given origin. A good joint fit does not necessarily imply a good fit of total out-migration, the quantity of importance for the origin population. For this reason, we give a second goodness-of-fit measure, binomial deviance for the total out-migration

$$\text{dev}_m = 2 \times N' \cdot \left[ (1 - {}_n\tilde{m}) \log \left( \frac{1 - {}_n\tilde{m}}{1 - {}_n m} \right) + {}_n\tilde{m} \log \left( \frac{{}_n\tilde{m}}{{}_n m} \right) \right]. \tag{29}$$

Low deviances can sometimes be achieved by an unrealistic profile, and for this reason we include the measure $\bar{P}$, equal to the sum of the percentage that each destination-specific schedule's profile differs from a reference profile

$$\bar{P} = \sum_{j=1}^{d} 100\left(1 - \frac{{}_n m'_{\text{ref}} \cdot {}_n m^j}{|{}_n m_{\text{ref}}||{}_n m^j|}\right), \tag{30}$$

where ${}_n m_{\text{ref}}$ is a reference schedule and $|v|$ is the absolute value of vector $v$. For reference schedule we used the P-TOPALS smoothed interstate probabilities from Dyrting (2020). When we assessed each fit's deficiencies, we focused on whether it over-smoothed the student peak (S), under-smoothed advanced ages (E), or had an implausible shape (X).

Table 2 shows that, for all origins and intervals, our generation–distribution approach gives realistic profiles, with the lowest or equal lowest value for $\bar{P}$. For 1-year probabilities it has the lowest value for $\text{dev}_m$ for seven of the eight states, and the lowest value for dev for five of the eight states, with the kernel regression (SA, ACT, and NT out-migration) or student MMS (NT out-migration) achieving a lower value with an unrealistic profile. For 5-year probabilities the three methods are more evenly matched. Our approach has the lowest value of $\text{dev}_m$ for five of the eight states, but kernel regression has the lowest value for dev for seven of the eight states. Our approach always has a lower value for dev compared to student MMS but has the lowest value for only one state (NSW).

## 7 Application to the 1-year/5-year problem

We previously described the 1-year/5-year census migration problem, which is hallmarked by differences in aggregate migration flows, and therefore derived probabilities, when comparing across time and space. Five-year probabilities are less than five times 1-year probabilities, the difference being due to multiple moves over the longer interval (Rees 1977). Unlike probabilities, migration ratios appear to be more stable across different intervals (Rogers et al. 2003) and the method developed here is useful for comparing them because they share a common representation in terms of implied 1-year migration ratios (see Sect. A.1).

Figures 5 and 6 show 1-year and 5-year implied migration ratios from New South Wales and the Northern Territory respectively using data from the 2016 Census. The complete set of 56 origin–destination ratios for all states and territories is given in Figures S-33 to S-40 of Online Resource 1. We see that the 5-year migration ratios implied by 5-year data have the same level as ratios from 1-year data and similar age profiles. This suggests that a crude method for converting 5-year probabilities to 1-year probabilities would be to focus on converting the generation component, keeping the distribution component fixed at its implied 1-year values.

## 8 Discussion and conclusion

This paper has extended the P-spline method to the problem of smoothing the migration ratio component of a generation–distribution representation of origin–destination-age-specific migration probabilities. Existing methods address this problem

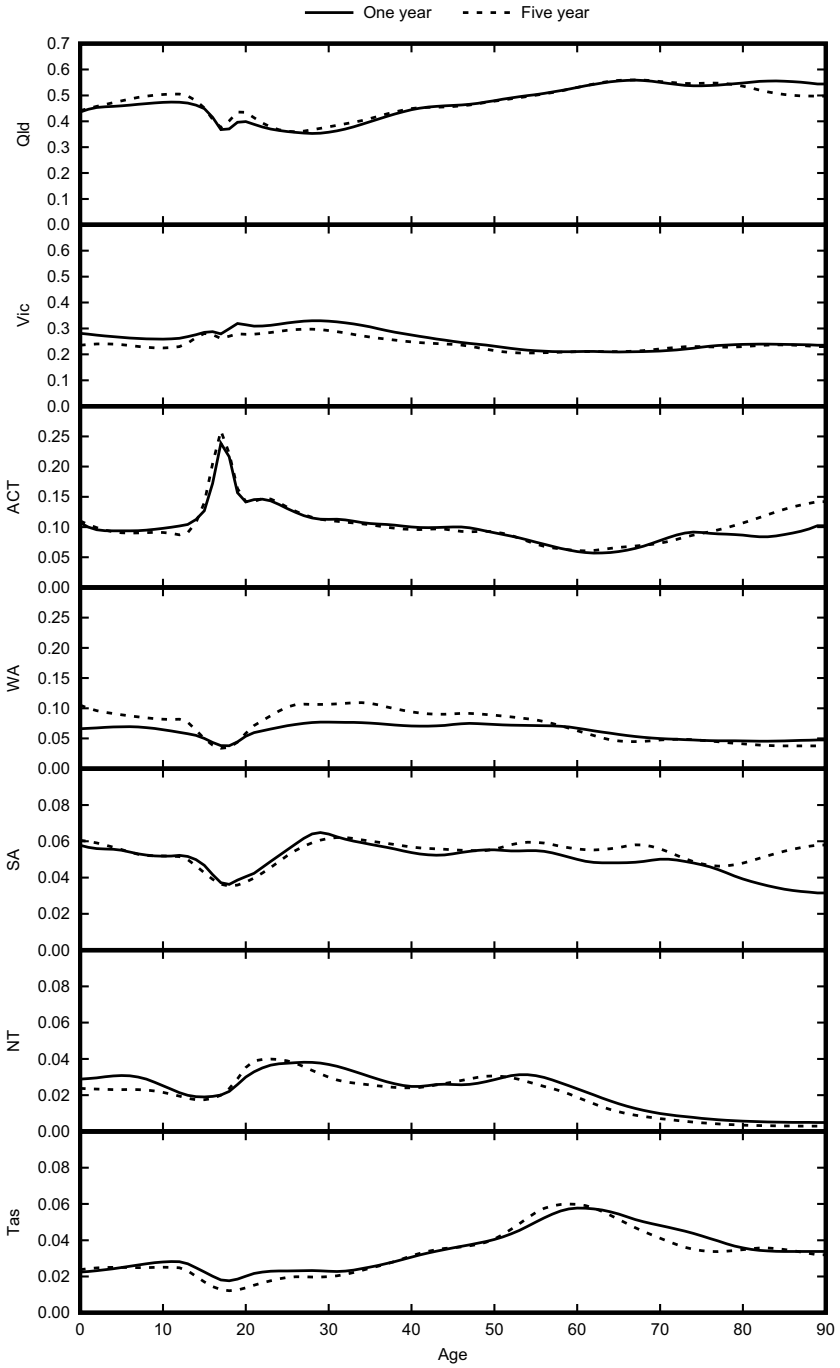


**Fig. 5** New South Wales interstate out-migration ratios 2016, two intervals. *Source*: Based on Australian Bureau of Statistics data

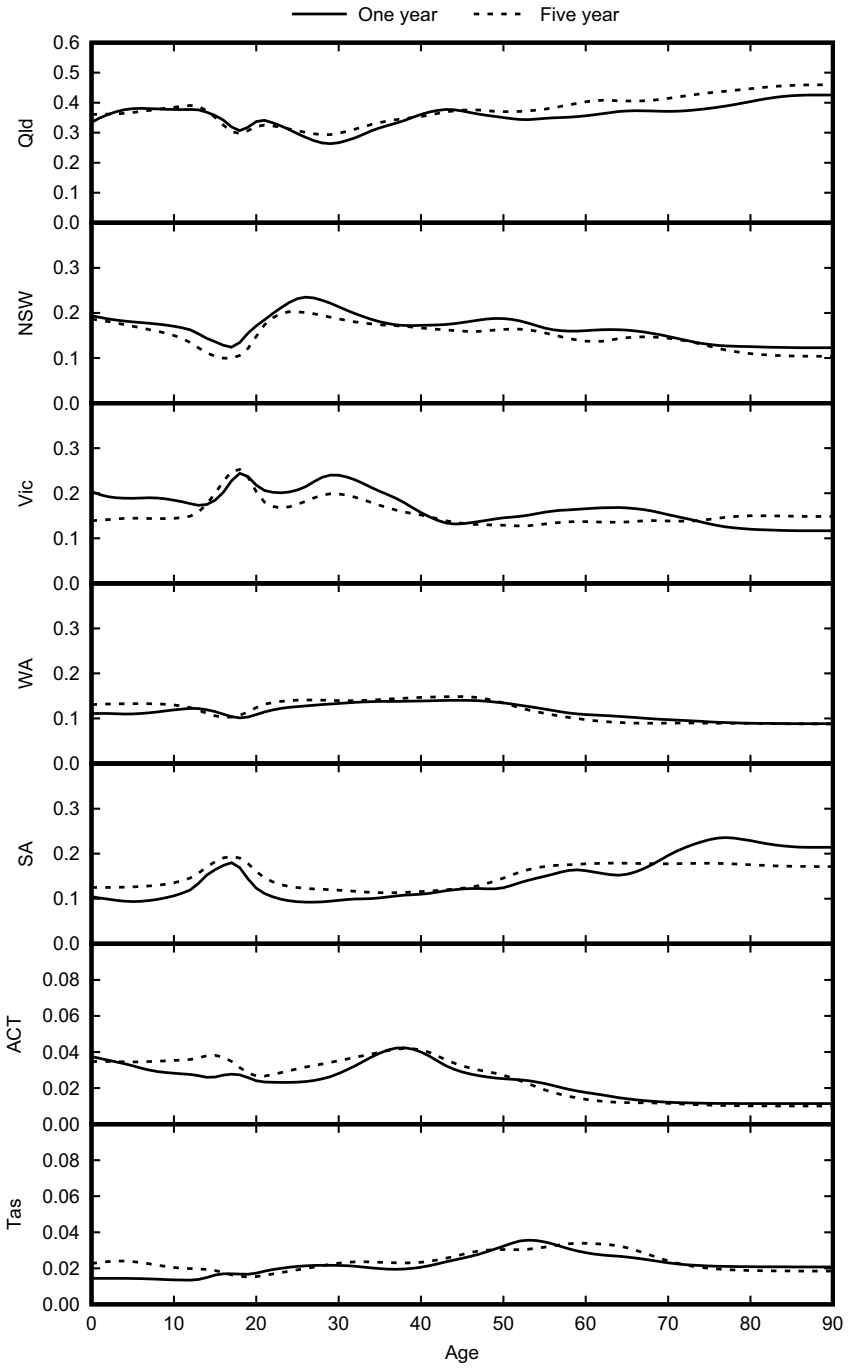


**Fig. 6** Northern Territory interstate out-migration ratios 2016, two intervals. *Source*: Based on Australian Bureau of Statistics data

through smoothing of bi-regional flows, primarily in-migration or out-migration with one destination–origin pairing (usually a State with the rest of the country). The contribution of our method is to provide a multi-regional approach to smoothing destination-migration profiles. The method has been implemented as an Excel add-in which is included in Online Resource 2. The potential downstream benefits from this method include the preparation of more accurate inputs for origin–destination-specific population projection models, and the construction of multi-regional life tables.

Using the example of Australian interstate migration, we have shown how P-splines can give an accurate fit to the migration ratio profile across high-curvature ages and a good treatment of sample noise both when the population at risk is low, such as advanced ages, and when the destination has a low conditional probability of migration. When combined with the use of P-TOPALS for smoothing the generation component we find that the P-spline method produced smooth origin–destination profiles that were both realistic and accurate, performing better than both kernel regression and student MMS for 1-year probabilities, but for 5-year probabilities the results for the three methods were more evenly matched. We used the method to directly compare 1-year migration ratios and implied 1-year ratios from 5-year data and found that they have the same level and similar age-specific shapes.

The framework we have adopted here, the generation–distribution decomposition of migration flows, is the same as used by Rogers et al. (2002), although our focus and tools are different. One difference is that we use single year of age data and smooth with P-splines while they used 5-year age groupings and smoothed using a log-linear model. The second difference is that we assume a complete set of data while they were interested in the problems posed by incomplete data, in particular the problem of repairing incomplete data by imposing age and spatial structures from an external source. The extent that the distribution component differs for 1-year and 5-year intervals is currently an open question. Liaw (1984) conjectured that most of the difference between one- and 5-year migration probabilities is due to the generation component, while Rogerson (1990) argues that the distribution component is also affected by the interval width. Rogers et al. (2003) concluded that the assumption of constant distribution component needed to be relaxed and some type of variation introduced through exogenous covariates. Our contribution to this debate has been to provide a representation of the distributional component in terms of implied 1-year ratios which enables 1-year and 5-year ratios to be directly compared.

Migration between the eight Australian states and territories is a good test case for our method because it spans two orders of magnitude of migration flow sizes from an average of 398.5 persons per single year of age of movers from New South Wales to Queensland over interval 2015–2016 to an average of 3.5 persons per single year of age from the Northern Territory to Tasmania over the same period. How would our method adapt to smaller spatial scales?, or more precisely, as the spatial scale is decreased is the method robust as the flow size decreases and manageable as the number of possible destinations increases? As the flow size ${}_nK$ decreases the penalty term in Eq. (15) will become more important. For the case of a linear penalty ($k = 1$) the B-spline weight will tend to a constant

$$\theta = \theta_0 \iota, \tag{31}$$

where $\iota$ is a vector of ones. Because B-splines form a partition of unity ($B \cdot \iota = 1$) the conditional ratio $a$ will, in this limit, tend to a value independent of age. The method therefore adaptively converges to the OPCS model described by Wilson and Bell (2004), where migrants are distributed to destinations in a fixed proportion regardless of age. Another aspect of applying the smoothing method at smaller spatial scales is the increase in the dimension of the multi-regional migration matrix, which scales as the square of the number of regions. The burden on the user of estimating, in one stage, the entire base period matrix has been cited as one of the challenges to practical implementation of a multi-regional population projection model (Wilson and Bell 2004). One of the strengths of our method is that it allows a multi-stage approach to estimating the matrix: for each origin estimate the out-migration rate, and sequentially for each destination, estimate the conditional migration ratio.

Assumption-setting, the processes of projecting the future trajectories of migration rates which are then used in population projections, while different from the estimation of current levels (which our method seeks to solve) is frequently connected to it (assumptions often being expressed as additive or multiplicative changes to the jump-off level) and presents similar challenges to the practitioner (e.g. how to manage dimensionality for small spatial scales). The framework we use enables users to handle the setting of assumptions by dividing them into two types (generation and distributional) which can be projected separately. Thus, for example, dynamism of out-flows can be modelled by making the total out-migration rate time-dependent, and dynamism of the distribution of flows can be modelled by making the migration ratios time-dependent.

The main strength of the P-spline method is its combination of flexibility in fitting the variety of age-specific profiles for migration ratios and ability to account for age-dependent sample noise. Another advantage is that it allows the practitioner to either use an automatic smoother such as the AICc condition or dial the level of smoothing manually for any origin–destination pair. A current limitation of the method is that it assumes data in the form of single years of age. Often census data on internal migration are published by age groups, commonly 5-year, and even when data are available by single year of age it is sometimes necessary to group it to mitigate the effects of age-heaping (Feeney 1979) or confidentialisation (Thompson et al. 2013). One area for further work, therefore, is to extend the P-spline approach to the distribution part when data is grouped. Another path for further investigation is to adapt the method for experimentally projecting the distribution component or more generally updating the distributional component conditional on partial information, possibly drawing on methods from Plane (1981).

## Appendix 1: Implied and conditional ratios

In this section we derive expressions for ratios and conditional ratios in terms of implied 1-year variables. The first step is to re-express the relationship between out-migration ${}_nm_x$ and implied 1-year probabilities $m_x$

$$ {}_nm_x = 1 - \prod_{x \le k < x+n} (1 - m_k), \tag{32} $$

in the form

$$ {}_nm_x = \sum_{x \le r < x+n} \left( \prod_{x \le k < r} (1 - m_k) \right) m_r. \tag{33} $$

Equation (33) is obviously true for $n = 1$. Assume it is true for some $n$. By Eq. (32)

$$ {}_{n+1}m_x = (1 - {}_nm_x) m_{x+n} + {}_nm_x. \tag{34} $$

Substituting expression (32) into the first term of the right-hand side of the above equation and expression (33) into the second term gives Eq. (33) for $n + 1$. The general result follows by induction. Equation (33) expresses the probability of out-migration from age $x$ to $x + n$ as the product of remaining for $x$ to $r$ (the term in parentheses) and then out-migrating from $r$ to $r + 1$.

### Implied ratios

By analogy we define implied ratios

$$ c^j = \begin{bmatrix} c_0^j \\ \vdots \\ c_\omega^j \end{bmatrix}, \tag{35} $$

through the expression for ${}_nm_x^j$,

$$ {}_nm_x^j = \sum_{x \le r < x+n} \left( \prod_{x \le k < r} (1 - m_k) \right) m_r c_r^j. \tag{36} $$

It follows from Eq. (27) that

$$ {}_nc^j = {}_nU \cdot c^j \tag{37} $$

where ${}_nU$ is the matrix with elements given by Eq. (13).

### Conditional ratios

The observed ratio ${}_n\tilde{a}^j$ of migrating to $j$ conditional on not migrating to destinations $1, \ldots, j-1$ is

$$ {}_n\tilde{a}^j := \frac{{}_nM^j}{{}_nK^j} = \frac{{}_n\tilde{c}^j}{\sum_{k=j}^d {}_n\tilde{c}^k}. \tag{38} $$

By definition ${}_n\tilde{a}^1 = {}_n\tilde{c}^1$ and ${}_n\tilde{a}^d = 1$. Equation (38) can be rearranged to give

$$ {}_n\tilde{c}^j = {}_n\tilde{s}^j \times {}_n\tilde{a}^j, \tag{39} $$

where

$$ {}_n\tilde{s}^j := \sum_{k=j}^{d} {}_n\tilde{c}^k = \prod_{1 \le k < j} (1 - {}_n\tilde{a}^k). \tag{40} $$

The last expression in Eq. (40) follows from the recurrence

$$ {}_n\tilde{s}^j = {}_n\tilde{s}^{j-1} - {}_n\tilde{c}^{j-1} = (1 - {}_n\tilde{a}^{j-1})_n\tilde{s}^{j-1}. \tag{41} $$

Substituting Eqs. (39) and (5) into the log likelihood function (4) we get Eq. (8).

#### Implied conditional ratios

Implied 1-year conditional ratios $a^j$ are related to implied ratios $c^j$ by Eq. (5) with $n = 1$, namely

$$ c^j = s^j \times a^j \tag{42} $$

where

$$ s^j = \prod_{1 \le k < j} (1 - a^k). \tag{43} $$

Substituting Eqs. (5) and (42) into Eq. (37) and rearranging we get Eq. (11) where

$$ {}_nT^j = \operatorname{diag}\left(\frac{1}{{}_ns^j}\right) \cdot {}_nU \cdot \operatorname{diag}\left(s^j\right). \tag{44} $$

The matrix ${}_nT^j$ is efficiently calculated using the recurrence (12).

## Appendix 2: Maximising the penalised likelihood function

The maximum of the function given by Eq. (15) satisfies equation

$$ \frac{\partial \mathscr{L}}{\partial \theta} = 0. \tag{45} $$

Taking the derivative we get the system of equations

$$ G' \cdot V \cdot \left({}_n\tilde{a} - {}_na\right) - \lambda D_k' \cdot D_k \cdot \theta = 0 \tag{46} $$

where

$$ V := \operatorname{diag}\left({}_nK\right). \tag{47} $$

and

$$G := \frac{1}{{}_n a(1 - {}_n a)} \frac{\partial_n a}{\partial \theta}. \tag{48}$$

Taking derivative of both sides of Eq. (11) and using the expression

$$\frac{\partial a}{\partial \theta} = a(1 - a)B \tag{49}$$

which follows from Eq. (14) we get the following expression for $G$

$$G = \text{diag}\left(\frac{1}{{}_n a(1 - {}_n a)}\right) \cdot {}_n T \cdot \text{diag}(a(1 - a)) \cdot B. \tag{50}$$

To solve Eq. (46) I use the approximations

$${}_n a(\theta) \approx {}_n a(\bar{\theta}) + {}_n a(\bar{\theta})\big(1 - {}_n a(\bar{\theta})\big) G \cdot (\theta - \bar{\theta}), \tag{51}$$

which when substituted into Eq. (46) gives the linear iteration Eq. (16).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00168-021-01051-4.

**Funding** This research was in part funded by the Northern Territory Department of Treasury and Finance.

**Availability of data and material** A selection of the data and examples of the smoothing method are available as an Excel spreadsheet from the corresponding author on request.

## Compliance with ethical standards

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Code availability** An implementation of the P-spline smoothing method as an Excel add-in is available from the corresponding author on request.



## References

ABS (2016) Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas. cat. no. 1270.0.55.001, Australian Bureau of Statistics

ABS (2019) Australian Demographic Statistics, Dec 2019. cat. no. 3101.0, Australian Bureau of Statistics
Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
Baffour B, Raymer J (2019) Estimating multiregional survivorship probabilities for sparse data: an application to immigrant populations in Australia, 1981–2011. Demogr Res 40(18):463–502
Bell M, Charles-Edwards E, Ueffing P, Stillwell J, Kupiszewski M, Kupiszewska D (2015) Internal migration and development: Comparing migration intensities around the world. Popul Dev Rev 41(1):33–58
Bell M, Charles-Edwards E, Bernard A, Ueffing P (2017) Global trends in internal migration. In: Champion T, Cooke T, Shuttleworth I (eds) Internal migration in the developed world: are we becoming less mobile?. International Population Studies. Routledge, London, pp 76–97
Bernard A, Bell M (2015) Smoothing internal migration age profiles for comparative research. Demogr Res 32(33):915–948
de Boor C (2001) A practical guide to Splines, Revised edn. Springer, New York
Camarda CG (2012) MortalitySmooth: an R package for smoothing Poisson counts with P-splines. J Stat Softw 50(1):1–24
Campbell PR (1996) Population projections for states by age, sex, race, and Hispanic origin: 1995 to 2025. PPL-47, U.S. Bureau of the Census, Population Division
Congdon P (2008) Models for migration schedules: a Bayesian perspective with applications to flows between Scotland and Wales. In: Raymer J, Willekens F (eds) International Migration In Europe: data, models and estimates. Wiley, Chichester, pp 193–205
Courgeau D (1973) Migrants and migrations. Population 28:95–128
Currie ID, Durbán M, Eilers PHC (2004) Smoothing and forecasting mortality rates. Stat Model 4(4):279–298
Dyrting S (2020) Smoothing migration intensities with P-TOPALS. Demogr Res 43(55):1607–1650
Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11(2):89–121
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London
Feeney G (1979) A technique for correcting age distributions for heaping in multiples of five. Asian Pac Census Forum 5(3):12–14
Gonzaga MR, Schmertmann CP (2016) Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. Revista Brasileira de Estudos de Populacao 33(3):629–652
Hachen DS (1988) The competing risks model. Soc Methods Res 17(1):21–54
Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76(2):297–307
Liaw KL (1984) Interpolation of transition matrices by the variable power method. Environ Plan A 16:917–925
Nash A (2020) Methodology used to produce the 2018-based subnational population projections for England. Office for National Statistics
Peters P, Taylor A, Carson DB, Brokensha H (2016) Sources of data for settlement level analyses in sparsely populated areas. In: Taylor A, Carson DB, Ensign PC, Huskey L, Rasmussen RO, Saxinger G (eds) Settlements at the Edge: remote human settlements in developed nations. Edward Elgar, Gloucester, pp 153–176
Plane DA (1981) Estimation of place-to-place migration flows from net migration totals: a minimum information approach. Int Reg Sci Rev 6(1):33–51
Rees P, Bell M, Duke-Williams O, Blake M (2000) Problems and solutions in the measurement of migration intensities: Australia and Britain compared. Popul Stud 2:207–222
Rees PH (1977) The measurement of migration, from census data and other sources. Environ Plan A 9(3):247–272
Rogers A (1975) Introduction to multiregional mathematical demography. Wiley, New York
Rogers A (1986) Population projections. In: Rogers A, Willekens FJ (eds) Migration and settlement. D. Reidel Publishing Company, Dordrecht, pp 211–263
Rogers A, Castro LJ (1981) Model migration schedules. Research Report RR-81-30. International Institute for Applied Systems, Laxenburg
Rogers A, Watkins J (1987) General versus elderly interstate migration and population redistribution in the United States. Res Aging 9(4):483–529
Rogers A, Raquillet R, Castro LJ (1978) Model migration schedules and their applications. Environ Plan A 10(5):475–502

Rogers A, Raymer J, Willekens F (2002) Capturing the age and spatial structures of migration. Environ Plan A 34:341–359
Rogers A, Raymer J, Newbold KB (2003) Reconciling and translating migration data collected over time intervals of differing widths. Ann Reg Sci 37(4):581–601
Rogers A, Little J, Raymer J (2010) The Indirect Estimation of Migration, 1st edn. Springer, Dordrecht
Rogerson PA (1990) Migration analysis using data with time intervals of differing widths. Papers Reg Sci Assoc 68:97–106
Ruppert D, Sheather SJ, Wand MP (1995) An effective bandwidth selector for local least squares regression. J Am Stat Assoc 90(432):257–1270
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Smith SK, Tayman J, Swanson DA (2013) A practitioner's guide to state and local population projections, demographic methods and population analysis, vol 37. Springer, Dordrecht
Thompson G, Broadfoot SJ, Elazar DJ (2013) Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. Paper presented at Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, October 28-30, 2013
Willekens F (2008) Models of migration: observations and judgements. In: Rayner J, Willekens F (eds) International Migration in Europe: data, models and estimates. Wiley, Chichester, pp 117–147
Willekens F (2016) Migration flows: measurement, analysis and modeling. In: White MJ (ed) International handbook of migration and population distribution, international handbooks of population, vol 6. Springer, Dordrecht, pp 225–241
Wilson T (2010) Model migration schedules incorporating student migration peaks. Demogr Res 23(8):191–222
Wilson T (2020) Modelling age patterns of internal migration at the highest ages. Spat Demogr 8:175–192
Wilson T, Bell M (2004) Comparative empirical evaluations of internal migration models in subnational population projections. J Popul Res 21(2):127–160
Wilson T, Terblanche W (2018) New estimates of Australia's centenarian population. Int J Popul Data Sci 3(1):1–10