

DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 32, ARTICLE 33, PAGES 915–948

PUBLISHED 12 MAY 2015

<http://www.demographic-research.org/Volumes/Vol32/33/>

DOI: 10.4054/DemRes.2015.32.33

Research Article

Smoothing internal migration age profiles for comparative research

Aude Bernard

Martin Bell

©2015 Aude Bernard & Martin Bell.

This open-access work is published under the terms of the Creative Commons Attribution NonCommercial License 2.0 Germany, which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/by-nc/2.0/de/>

Table of Contents

1	Introduction	916
2	Model migration schedules	918
3	Alternative non-parametric models	920
3.1	Cubic b-spline	920
3.2	Kernel regression	921
4	Data and methods	922
5	Evaluation of smoothing methods	925
5.1	Standard MMS versus cubic spline and kernel regression	925
5.2	Student MMS versus cubic spline and kernel regression	934
6	Research application	936
6.1	Comparison across countries	936
6.2	Comparison over time	939
7	Conclusion	941
	References	944

Smoothing internal migration age profiles for comparative research

Aude Bernard¹

Martin Bell²

Abstract

BACKGROUND

Age patterns are a key dimension to compare migration between countries and over time. Comparative metrics can be reliably computed only if data capture the underlying age distribution of migration. Model schedules, the prevailing smoothing method, fit a composite exponential function, but are sensitive to function selection and initial parameter setting. Although non-parametric alternatives exist, their performance is yet to be established.

OBJECTIVE

We compare cubic splines and kernel regressions against model schedules by assessing which method provides an accurate representation of the age profile and best performs on metrics for comparing aggregate age patterns.

METHOD

We use full population microdata for Chile to perform 1,000 Monte-Carlo simulations for nine sample sizes and two spatial scales. We use residual and graphic analysis to assess model performance on the age and intensity at which migration peaks and the evolution of migration age patterns.

RESULTS

Model schedules generate a better fit when (1) the expected distribution of the age profile is known a priori, (2) the pre-determined shape of the model schedule adequately describes the true age distribution, and (3) the component curves and initial parameter values can be correctly set. When any of these conditions is not met, kernel regressions and cubic splines offer more reliable alternatives.

CONCLUSION

Smoothing models should be selected according to research aims, age profile characteristics, and sample size. Kernel regressions and cubic splines enable a precise

¹ University of Queensland, Australia. E-Mail: a.bernard@uq.edu.au.

² University of Queensland, Australia. E-Mail: martin.bell@uq.edu.au.

representation of aggregate migration age profiles for most sample sizes, without requiring parameter setting or imposing a pre-determined distribution, and therefore facilitate objective comparison.

1. Introduction

Internal migration is an age-selective process, with young adults being the most mobile group. Migration intensities typically decrease from birth to the teenage years before peaking at young adult ages. They then decline steadily with increasing age, sometimes rising again around the age of retirement. Rogers and Castro (1981) have demonstrated that this broad age profile is replicated across a number of countries and at various spatial scales. Subsequent analysts have proposed summary indicators of the age profile that can be used to make comparisons across countries and over time (Bell et al. 2002; Bernard et al. 2014a) and to examine the association of the migration age profiles with other demographic processes such as life-course age patterns (Bernard et al. 2014b).

Since its introduction by Rogers et al. (1978), the model migration schedule has been widely adopted as the main method to smooth migration age profiles. As a composite exponential function, the model schedule constrains migration to follow a prototypical shape founded on the theoretical link between migration and life-course transitions. While model schedules are based on accumulated evidence primarily from contemporary Europe and the United States, they constrain migration age patterns to follow a standard predicted shape that may not describe migration age patterns accurately in other regions or in historical populations. Furthermore, model schedules face limitations related to their estimation (Bernard et al. 2014a), including the difficulty of selecting an optimal set of component curves specific to each age profile (Rees et al. 2000), the instability of parameter estimates (Congdon 1993), and the sensitivity of estimates to initial parameter values (Rogers et al. 2005). Researchers have therefore to employ a trial-and-error approach to decide on the set of component curves and initial parameter values that will yield the best fit, which directly influences the shape of the estimated curve and the value of the estimated parameters. This, in turn, undermines the reliability of these parameters for comparative analysis.

Demographers have recourse to a range of other statistical methods for data smoothing, including non-parametric models such as cubic splines and kernel regressions, which have been widely used to smooth fertility (Moguerza et al. 2010) and mortality age profiles (Peristera and Kostaki 2005). Non-parametric models have the advantage of avoiding the imposition of a pre-determined shape (Pagan and Ullah 1999) and are also easier to implement using automated processes free of sensitivity to

subjective assumptions (Wand and Jones 1995). As a result of the widespread adoption of model schedules, there appears to have been no previous attempt to systematically assess the different methods available to smooth migration age profiles.

The age and intensity at which migration peaks have widely been used to characterise and compare migration age patterns across countries, either alongside other metrics (Rogers and Castro 1981) or as the main or sole summary measures (Bell et al. 2002; Bell and Muhidin 2009; Bernard et al. 2014a; Bracken and Bates 1983; Rees et al. 2000). The choice of a particular model and its specification affect the shape of the fitted curve, which in turn influences estimates of the age and intensity at peak. Inappropriate model selection or incorrect model misspecifications are therefore likely to result in incorrect inferences when comparing countries on these measures or when examining the evolution of migration age patterns within a country. Comparative analysis of migration calls for smoothing methods that preserve the overall distribution shape of the age profile and retain discriminating features.

This paper seeks to evaluate and compare the strengths and limitations of cubic splines, kernel regressions, and model schedules for smoothing migration age profiles at a range of sample sizes. Our particular concern is to assess the performance of these three models on the key indicators used for comparing aggregate migration age patterns across countries and over time, namely the age at which migration peaks and the intensity at the peak (Bernard et al. 2014a). This work forms part of the IMAGE project (Internal Migration Around the Globe, <http://www.gpem.uq.edu.au/image>), an international programme of collaborative research. Its aim is to develop and implement a set of rigorous statistical indicators to measure several dimensions of internal migration, including age, which can be used to make comparisons between countries and over time. The paper is organised as follows. Section 2 reviews model migration schedules and describes the issues associated with their estimation. Section 3 presents cubic splines and kernel regressions and identifies their strengths and limitations. Section 4 describes the procedure to estimate the three models and presents the data. We focus on aggregate migration, which captures all moves within a country, irrespective of distance or direction, as the most appropriate basis for cross-national comparisons (Bell et al. 2002). Using full population microdata from the 2002 Chilean census, Section 5 performs 1,000 Monte-Carlo simulations for nine sample sizes and two spatial scales, and generates 18,000 random samples. We then evaluate the goodness-of-fit of each model based on residual analysis and graphical validation. In section 6, we illustrate the practical significance of using different smoothing techniques by examining the performance of each model on the age and intensity at which migration peaks in Argentina, Chile, France, and Greece and on the evolution of migration age patterns in Chile since the 1980s. Section 7 concludes the paper and discusses other possible research applications.

2. Model migration schedules

Rogers et al. (1978) were the first to draw on the persistent shape of migration age profiles to establish a mathematical model that summarises and codifies such regularities. Denoted a model migration schedule (MMS), this composite exponential function comprises a childhood curve, a labour force curve, a constant, and in some instances a retirement curve. The reduced form of the model schedule is the sum of three component functions and comprises seven parameters. Algebraically, the migration intensity m at age x is expressed as:

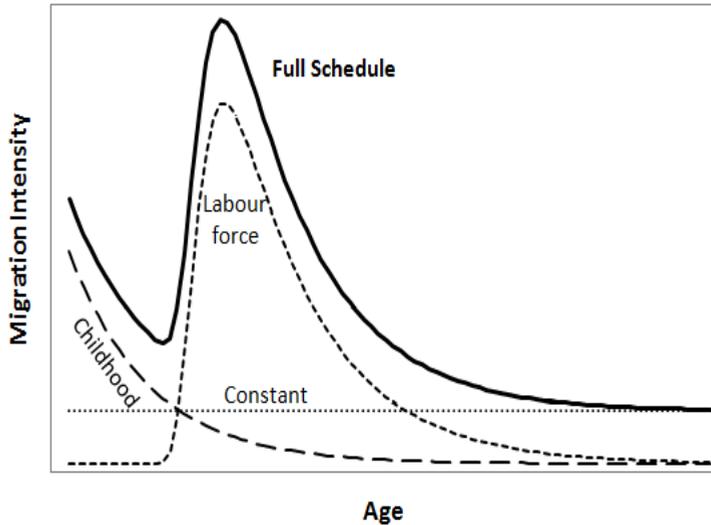
$$\hat{m}(x) = a_1 \exp(-\alpha_1 x) + a_2 \exp\{-\alpha_2 (x - \mu_2) - \exp[-\lambda_2(x - \mu_2)]\} + c \quad (1)$$

the parameters of which are conventionally defined as follows:

- a_1 = height of the childhood curve
- α_1 = rate of descent of the childhood curve
- a_2 = height of the labour force curve
- λ_2 = rate of ascent of the labour force curve
- α_2 = rate of descent of the labour force curve
- μ_2 = position of the labour force curve on the age axis
- c = constant

The first term of the model schedule is associated with children and teenagers. It is a negative exponential function starting from an initial maximum value of a_1 , and then declining at a rate of α_1 thereafter. The second term is a double exponential function with four parameters that describe the level, shape, and position of the labour force curve. The height of the curve is defined by a_2 and the location of the curve on the age axis is captured by μ_2 . The rates of ascent and descent are respectively characterised by λ_2 and α_2 . Finally, model schedules include a constant c , which defines a base level of migration across all ages. Figure 1 illustrates the three constituent curves of the reduced form of the model migration schedule. A retirement curve can be added to the reduced form of model schedule, typically reflecting out-migration flows from metropolitan areas to warm, coastal, and high-amenity areas (Rogers 1988). This model was later extended through the addition of an elderly upward slope (Rogers and Watkins 1987) and a student curve (Wilson 2010).

Figure 1: Reduced form of the model migration schedule and its component curves



Source: Based on Rogers and Castro, 1981

In a recent review of model migration schedules, Bernard et al. (2014a) identified two main limitations related to the estimation of model schedules. The first limitation bears upon the difficulty of selecting the appropriate number of component curves. A total of eight different combinations of component curve are possible depending on the functions included in the model, and selecting an appropriate set of constituent curves requires a priori knowledge of the expected age distribution. The problem is exacerbated by the fact that the choice of component curves included in the model affects the values of the parameters estimated, and hence the shape of the curve. The second limitation arises from the sensitivity of the fitted curve to the initial parameter values. Fitting model schedules to observed age-specific migration intensities requires non-linear curve fitting programs (Rogers and Little 1994; Rogers and Raymer 1999; Rogers et al. 2010; Wilson 2010). The algorithm is seeded with an initial set of user-specified parameters and produces a revised set of 'optimum' parameter estimates by iteratively substituting alternative values until pre-determined convergence criteria are met. Changes in the initial parameter values can result in widely varying fitted curves (Rogers et al. 2005). Limited guidelines exist around how to choose a 'good' set of initial parameter values and these are usually drawn from previous studies (Congdon

2008). However, typical values are only available for a handful of countries, principally the US and the UK, and no guidelines exist on how to best select initial values. Researchers have to make informed decisions, employing trial-and-error approaches to decide which component curves and which initial parameter values will yield the best fit. However, if the initial parameter values set by users significantly depart from the shape of the observed age profile, the residuals generated over many iterations may exceed an acceptable threshold for convergence, with the result that the model may fail to find a solution.

3. Alternative non-parametric models

3.1 Cubic b-spline

Demographers were introduced to cubic splines by Shryock and Siegel (1975), and McNeil et al. (1977) popularised their use to interpolate and smooth demographic data. Since then, cubic splines have been widely used to interpolate migration data based on five-year age groupings (Rogers and Castro 1981; Rogers et al. 2010), and smooth migration data classified by single years of age (Castro and Rogers 1983). Compared with model schedules, cubic splines impose fewer restrictions on the shape of the regression curve. They are linear combinations of polynomial pieces of degree 3 connected at knots, with continuous second derivatives. In this paper we have specified b-splines as they provide more stable estimates (Schoenberg 1969). For any given set of user-specified knots s , a b-spline can be expressed as:

$$B(x; s_1, \dots, s_{k+2}) = (k + 1) \sum_{j=1}^{k+2} \left[\prod_{1 \leq m \leq k+2, m \neq j} (s_m - s_j) \right]^{-1} P_k(x; s_j) \quad (2)$$

where k is the degree of the polynomial, which is set to 3 for cubic splines, and where P_k can be expressed as:

$$P_k(x; s) = \begin{cases} (x - s)^k, & x \geq s \\ 0, & x < s \end{cases} \quad (3)$$

Cubic splines are likely to deliver a better fit than model schedules because they use local information more intensively and may have a higher number of parameters. They are, however, highly sensitive to the number of knots and their location (Carter and Signorino 2010). The number of knots determines the degree of smoothness of the

curve, while their location establishes the points at which the relationship between age and migration may change significantly. Knots are usually selected by examination of two-dimensional scatter plots (Ruppert et al. 2003) and model fit criteria such as F-tests (Beck et al. 1998). Theoretical and empirical knowledge, such as the location of modes, can provide additional guidance for knot selection. Equidistant knots can be used but their small and discrete number allows only limited control over smoothness and fit. In applications to migration, the number of knots has generally been fixed and their locations distributed at five-year intervals (Rogers and Castro 1981). A trial-and-error approach to knot selection can improve the overall fit of the model, by ensuring that the regression curve captures the ages at which the relationship between age and migration changes. Adaptive data-driven methods for knot-selection have recently been proposed in the engineering literature (Biller and Farhrmeir 2001; Sharef et al. 2010), but remain computationally intensive.

3.2 Kernel regression

Kernel regression is a popular non-parametric model to estimate the conditional expectation of a random variable, without any distributional assumptions (Wand and Jones 1995). In fact, kernel regression has been used to smooth age-specific fertility rates (Moguerza et al. 2010) and mortality rates (Peristera and Kostaki 2005); however, it does not appear to have been applied to the task of smoothing migration data disaggregated by age. Computed between ages 5 to 90, a kernel regression estimates migration as a continuous function of age as follows:

$$\hat{m}(x) = \frac{\sum_{a=5}^{90} \varphi \left[\frac{x-a}{h} \right] m_a}{\sum_{a=5}^{90} \varphi \left[\frac{x-a}{h} \right]} \quad (4)$$

where φ is a normal density function. It is a decreasing function of $|x - a|$, so that it allocates more weight to the observed migration intensities close to age x and less weight to observed migration intensities further from age x . The choice of a bandwidth $h > 0$, or smoothing parameter, is critical. Large bandwidths tend to reduce the variance by smoothing over a large number of data points but risk failing to capture essential features or even distorting the underlying structure of the distribution, while small bandwidths maintain greater nuance in the shape of the distribution but are less effective in reducing noise. In this paper we use the default ROT bandwidth selection in Stata's `lpolys` function. The algorithm can be found in Stata User's Guide (StataCorp 2013). There is general agreement that the choice of kernel function is not critical for

the performance of the regression, but that the choice of bandwidth is (Marron and Nolan 1988). We use a Gaussian function as the reference distribution, as recommended by Härdle et al. (2004). For validation, we compared its performance against Epanechnikov and Quartic functions and found that it generates smaller sums of squared residuals. Differences in residuals between Gaussian, Epanechnikov, and Quartic functions are, however, minimal.

4. Data and methods

In applying a smoothing technique, researchers are seeking to reduce noise from sample variance while retaining the true underlying age pattern of migration (Goodall 1990). The task is particularly challenging with small samples, which can cause substantial irregularities. Thus we evaluate the performance of three estimators - model migration schedules, kernel regressions, and cubic splines – at a range of sample sizes. To that end, we use full population microdata from the 2002 Chilean census, which we take to represent the true age distribution of migration within Chile, and draw random samples with replacement of nine different sizes ranging from 25,000 to 10 million observations. To minimise noise and increase the reliability of the estimates, we use a Monte-Carlo simulation approach and repeat the above process 1,000 times at two spatial scales, reflecting migration between municipalities and between provinces, and generate a total of 18,000 samples.

Migration data in Chile are recorded as transitions based on comparing place of residence at the time of the census to that five years earlier. Observed migration data are therefore available only for individuals aged five and over on census night³. We compute aggregate migration between 335 municipalities and between 54 provinces. Aggregate migration refers to all moves, irrespective of distance or direction, and is widely regarded as the most appropriate basis for cross-national comparisons (Bell et al. 2002). Figure 2 represents the observed age profiles for the entire population aged 5 to 90 years⁴ drawn from the full population sample and from the first 10 randomly selected samples of size $n=25,000$, $n=250,000$, and $n=2$ million.

³ Migration age patterns of children below age 5 could be indirectly estimated by using birthplace-specific population stock data of the 0-to-4-year olds (Rogers and Jordan 2004). In the case of Chile, this technique would require access to the municipality of birth of the 0-to-4-year olds, which the authors did not hold at the time of the analysis. In any event, the migration behaviour of the birth cohort is peripheral to our central focus in this paper, which is the cross-national comparison of migration using key parameters which capture migration among young adults.

⁴ Single years above age 90 comprise less than 6,500 people in Chile, which is too low to ensure that each single year of age comprises enough observations to randomly select samples beyond age 90.

The full population sample shows that the shape of the curve follows the general age pattern of migration, characterised by a peak at young adult ages followed by a decline in migration intensities thereafter (Rogers and Castro 1981), and a slight increase late in life in response to moves to institutions (Rogers and Watkins 1987). The profile, however, departs from the standard migration age pattern in showing an increase in migration intensity in the late teenage years, suggesting the presence of movements for higher education, which are not captured by standard model migration schedules. Turning to Monte-Carlo simulations, using samples of $n=25,000$ does not reveal the presence of a student curve, but does show a marked increase in migration intensity at very old ages, probably caused by very small sample sizes at these ages. The student curve is distinguishable for samples with 250,000 observations or more and is most prominent for flows between municipalities.

Figure 2: Five-year migration intensity by single year, selected sample sizes

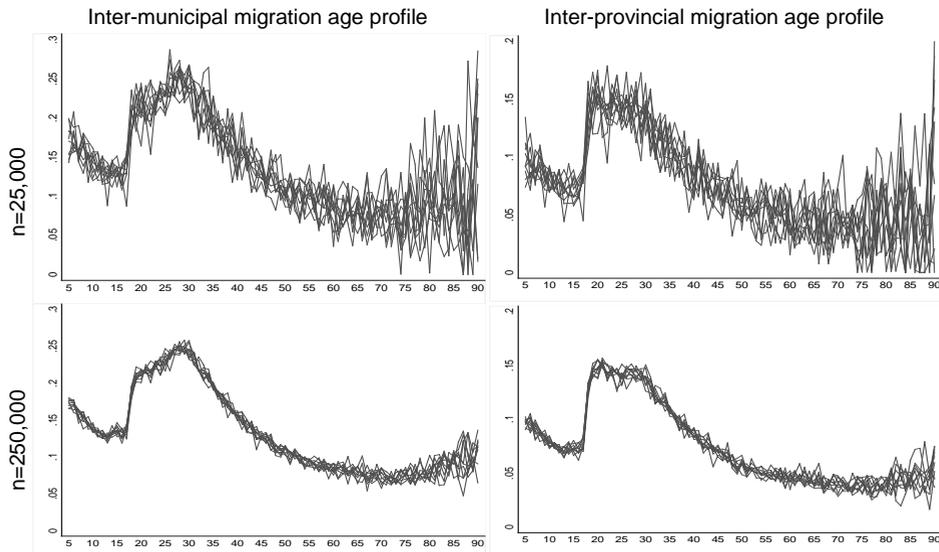
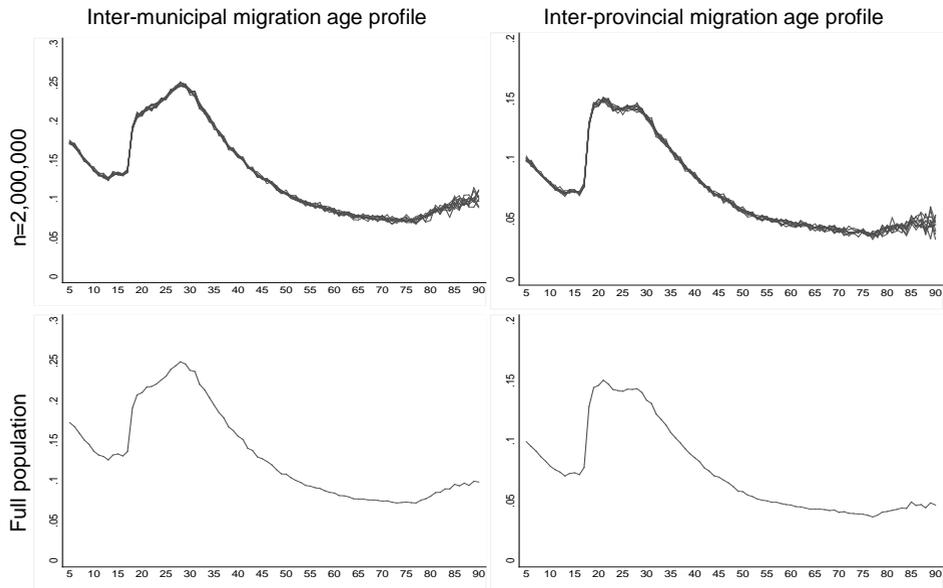


Figure 2: (Continued)



These results show that irregularities caused by small sample sizes can be problematic for the estimation of model migration schedules because there is no reliable basis on which to select an appropriate set of constituent curves and to set optimal initial parameter values. This is a very challenging task without a priori knowledge of the expected age distribution of migration. To illustrate common instances where researchers have limited information on a particular age profile, we first attempt to fit the Chilean data with standard model schedules, which include an upward elderly slope but no student curve, as described in Rogers and Watkins (1987). We denote this as the “standard MMS” and fit it using a version of the MATLAB program written by Rogers et al. (2010), amended to include an elderly slope. We compare its performance to cubic splines and kernel regressions fitted in Stata, which is more commonly used by demographers than MATLAB. To cover the whole age range we specify a cubic spline spanning 20 knots. We use one knot every three years between ages 15 and 30 to capture highly age-specific patterns at young adult ages, and one knot every five years at other ages. For kernel regressions we specify a Gaussian kernel function whose bandwidth is selected through an automated data-driven selection method according to Stata’s *lpoly* function (StataCorp 2013). The result is a bandwidth ranging from 1.71 to 1.94 for migration between municipalities, and from 1.69 to 2.34 for migration between

provinces. To take into account the increase in migration intensity in the early twenties observed in large samples we then add a student curve to the model migration schedule, which we denote “student MMS”. We do so for samples with at least 250,000 observations, which is the sample size from which the presence of a student curve can be observed in Chile, as shown in Figure 2. Table 1 describes each model in summary form, indicates the degrees of freedom for estimations between ages 5 to 90, and specifies the implementation requirements of each model.

Table 1: Model descriptions, degrees of freedom, and implementation requirements

Model	Description	Degrees of Freedom	Implementation
Model Migration Schedule	Standard MMS with four component curves as described in Rogers and Watkins (1987)	9	MATLAB program (Rogers et al. 2010) amended to include an upward elderly slope and a student curve for the student MMS.
	Standard MMS extended to include a student curve as described in Wilson (2010)	13	Users to specify component curves and set initial parameter values.
Kernel Regression	Gaussian-weighted average of rates at neighbouring ages as described in equation (4)	16.0 to 29.5*	Stata (<i>lpoly</i> function). No user specification required.
Cubic b-Spline	Piecewise cubic function with continuous second derivatives and knots at fixed values as described in the text and in equation (2)	18**	Stata (<i>bspline</i> function). Users to specific the number and location of knots.

*the effective number of parameters if given by the trace of the smoothing matrix

$$DF(h) = \sum_{a=5}^{90} \left\{ \varphi(0) / \sum_{z=5}^{90} \varphi \left[\frac{z-a}{h} \right] \right\} \text{ (Hastie and Tibshirani 1990)}$$

** While the b-spline syntax requires the inclusion of end-point knots at ages 5 and 90, they do not add to the number of parameters

5. Evaluation of smoothing methods

5.1 Standard MMS versus cubic spline and kernel regression

To estimate the performance of each smoother, we computed residuals at age a as the difference between the true migration intensity for the full population (m_a) and the fitted migration intensity (\hat{m}_{as}) at age a for sample size s . For each smoothed age

profile we calculated the sum of squared residuals (SSR) across ages 5 to 90. We then computed average SSR (*AvgSSR*) across the 1,000 random population samples for each model and sample size, as shown in equation (5):

$$AvgSSR = \frac{1}{S} \sum_s \sum_a (\hat{m}_{as} - m_a)^2 \quad (5)$$

We then decomposed *AvgSSR* into squared bias and variance components for each method and sample size as follows:

$$\begin{aligned} AvgSSR &= \frac{1}{S} \sum_s \sum_a (\hat{m}_{as} - m_a)^2 \\ &= \frac{1}{S} \sum_s \sum_a [(\hat{m}_{as} - \bar{m}_a) + (\bar{m}_a - m_a)]^2 \\ &= \sum_a \left[\frac{1}{S} \sum_s (\hat{m}_{as} - \bar{m}_a)^2 \right] + \sum_a (\bar{m}_a - m_a)^2 \\ &= AvgVariance + Avg Squared Bias \end{aligned} \quad (6)$$

where $\bar{m}_a = \frac{1}{S} \sum_a \hat{m}_{as}$ is the average fitted rate at age a .

Figures 3a and 3b represent *AvgSSR* by model and sample size for migration between municipalities and provinces. Because very small samples are skewed toward much larger *AvgSSR*, the results are displayed on a logarithmic scale. They show that for all models *AvgSSR* decreases with increasing sample size until $n=500,000$, and stabilises thereafter. The reduction in residuals is less pronounced for standard MMSs, which generate consistently high *AvgSSR*. While the performance of standard MMSs is comparable to that of kernel regressions for very small samples, it delivers the worst fit for samples greater than 250,000 observations. This is because it does not separately identify the peak that occurs at the time of entry into higher education. It constrains migration intensity to follow constant rates of ascent and descent around the peak, and therefore locates the estimated peak value between the student peak and the labour force peak. This can be seen in Figure 4, which displays average fitted inter-provincial migration age profiles for selected sample sizes.

Figure 3a: Log of *AvgSSR* by sample size, migration between municipalities

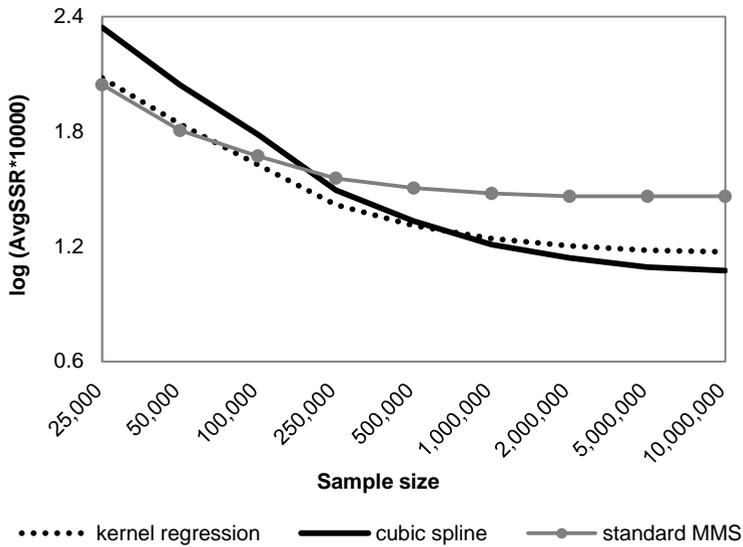


Figure 3b: Log of *AvgSSR* by sample size, migration between provinces

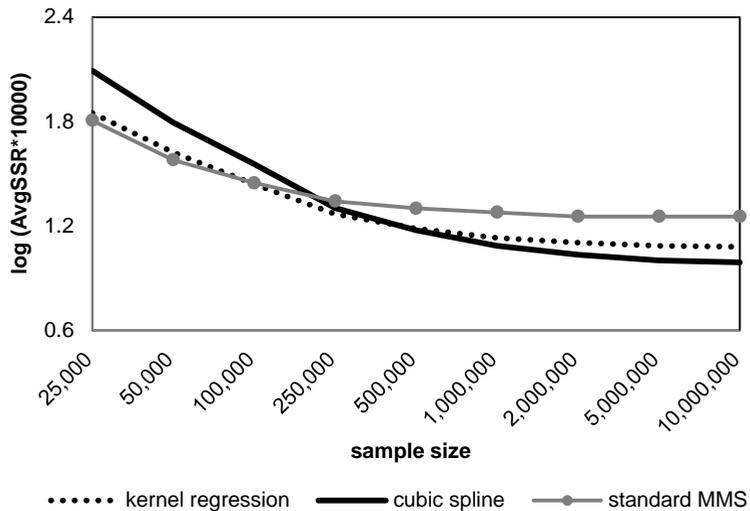
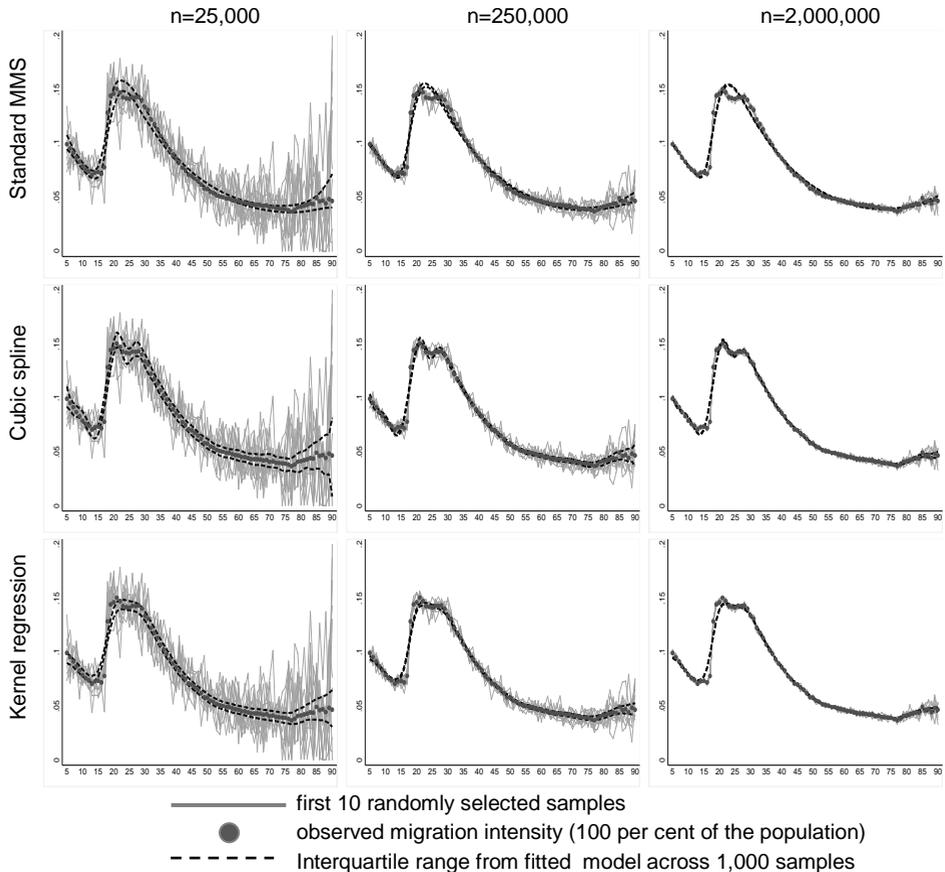


Figure 4: Inter-provincial migration age profiles by model, selected sample sizes



Cubic splines generate *AvgSSR* equivalent to kernel regressions for samples greater than $n=250,000$, but provide the worst fit of all models for smaller samples. Poor fit for very small sample sizes is due to the high sensitivity of cubic splines to knot location (Carter and Signorino 2010). Figure 4 shows that, while cubic splines provide a good fit for $n=250,000$ and $n=2$ million samples, they result in a poor fit for $n=25,000$ samples, where knots appear to capture noise rather than the true distribution beyond age 80. The interquartile range of fitted values is significantly larger than that of other models at older ages.

To further explore model performance, we decomposed *SSR* into squared bias and variance components, as defined in equation (6). Results are shown in Tables 2a and 2b and indicate that squared bias varies very little with sample size, whereas variance increases as the number of observations diminishes. This is particularly true for cubic splines, which exhibit significantly higher variance for samples smaller than $n=250,000$. This suggests that they should not be used to smooth very noisy samples. Turning to squared bias, cubic splines generate the smallest average squared bias for all sample sizes, while standard MMSs consistently exhibit the highest values. These findings suggest that kernel regressions may provide a good compromise between squared bias and variance for most sample sizes, and thus provide consistently more reliable estimates.

Table 2a: Decomposition of *AvgSSR* into squared bias and variance components, migration between municipalities

	Sample size	kernel regression	cubic spline	standard MMS
AvgSSR *10 ⁴	25,000	120	221	111
	50,000	69	111	64
	100,000	42	61	47
	250,000	26	31	36
	500,000	20	22	32
	1,000,000	17	16	30
	2,000,000	16	14	29
	5,000,000	15	12	29
	10,000,000	15	12	29
Average squared bias *10 ⁴	25,000	20	15	35
	50,000	18	13	29
	100,000	16	12	29
	250,000	15	12	29
	500,000	15	12	29
	1,000,000	15	12	28
	2,000,000	15	12	28
	5,000,000	15	12	29
	10,000,000	15	12	29
Average variance *10 ⁴	25,000	100	206	76
	50,000	52	98	35
	100,000	26	49	18
	250,000	11	19	7
	500,000	5	10	3
	1,000,000	2	4	2
	2,000,000	1	2	1
	5,000,000	0	1	0
	10,000,000	0	0	0

Table 2b: Decomposition of *AvgSSR* into squared bias and variance components, migration between provinces

	Sample size	kernel regression	cubic spline	standard MMS
AvgSSR	25,000	71	123	64
*10 ⁴	50,000	42	62	38
	100,000	28	36	28
	250,000	19	20	22
	500,000	15	15	20
	1,000,000	14	12	19
	2,000,000	13	11	18
	5,000,000	12	10	18
	10,000,000	12	10	18
Average squared bias	25,000	17	11	17
*10 ⁴	50,000	15	11	16
	100,000	13	10	17
	250,000	13	10	17
	500,000	12	10	18
	1,000,000	12	10	18
	2,000,000	12	10	18
	5,000,000	12	10	18
	10,000,000	12	10	18
Average variance	25,000	54	112	47
*10 ⁴	50,000	27	52	22
	100,000	14	26	11
	250,000	6	10	5
	500,000	3	5	2
	1,000,000	1	3	1
	2,000,000	1	1	0
	5,000,000	0	0	0
	10,000,000	0	0	0

We now explore model performance by examining the age distribution of average residuals, using $n=250,000$ samples by way of illustration. Figures 5a and 5b show a substantial increase in residuals at young adult ages for both inter-municipal and inter-provincial migration. A rapid increase in migration intensity can result in a disjuncture in the rate plots (Congdon 2008), and thus all models exhibit a rapid increase in residuals at ages 15 to 17, followed by a sharp decrease at ages 18 to 19, which correspond to the transition between the childhood and the labour force curves. Figure 4 shows that the high-curvature portion of the migration age profile, in the late teens to the mid-twenties, has the biggest fitting problem for all smoothers, even for larger samples, but is especially problematic for kernel regressions. While kernel regressions

and cubic splines exhibit consistently low residuals beyond the early twenties, standard MMSs display persistently high residuals until the mid-thirties, because they do not separately identify the student peak and the labour force peak, and produce a poor fit at young adult ages for all sample sizes.

Figure 5a: Age distribution of average residuals, inter-municipal migration

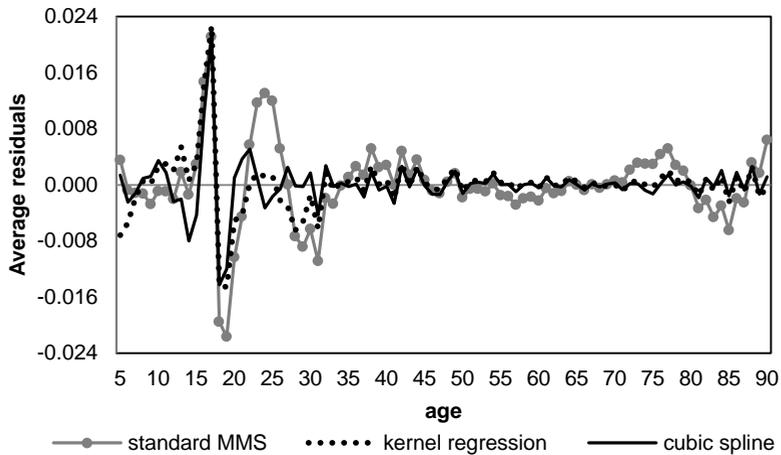
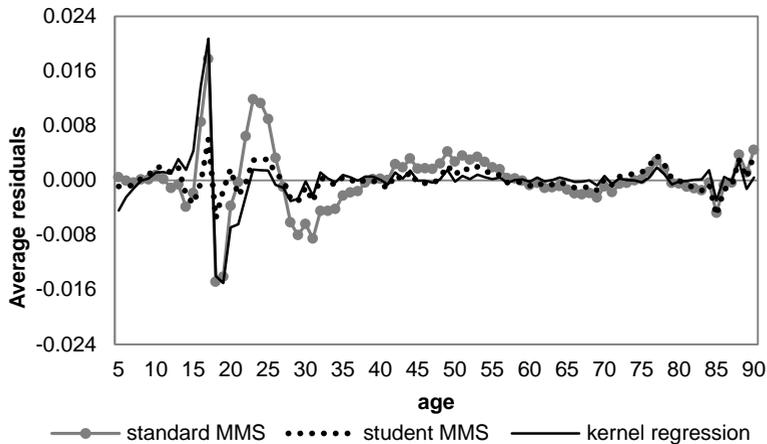


Figure 5b: Age distribution of average residuals, inter-provincial migration



Note: samples n=250,000

To further examine the age patterns of bias, we analysed first-order autocorrelation of average residuals by computing Pearson correlation coefficients between average residuals at age a and age $a + 1$ as defined in equation (7).

$$\rho = [(\bar{m}_a - m_a), (\bar{m}_{a+1} - m_{a+1})] \quad (7)$$

The results are reported by sample size and model in Figures 6a and 6b. Standard MMSs show a positive autocorrelation ($\rho > 0.50$) for all sample sizes, whereas cubic splines show no evidence of autocorrelation ($\rho < 0.15$). Residuals generated by kernel regressions vary with sample size, $0.34 < \rho < 0.47$, and show signs of autocorrelation for samples with less than 100,000 observations. These results confirm that fitted age profiles are more likely to sequentially depart from the true distribution when using standard MMSs because they constrain migration age patterns to follow a proto-typical distribution that does not systematically describe observed patterns.

Figure 6a: First-degree autocorrelation of average residuals, inter-municipal migration

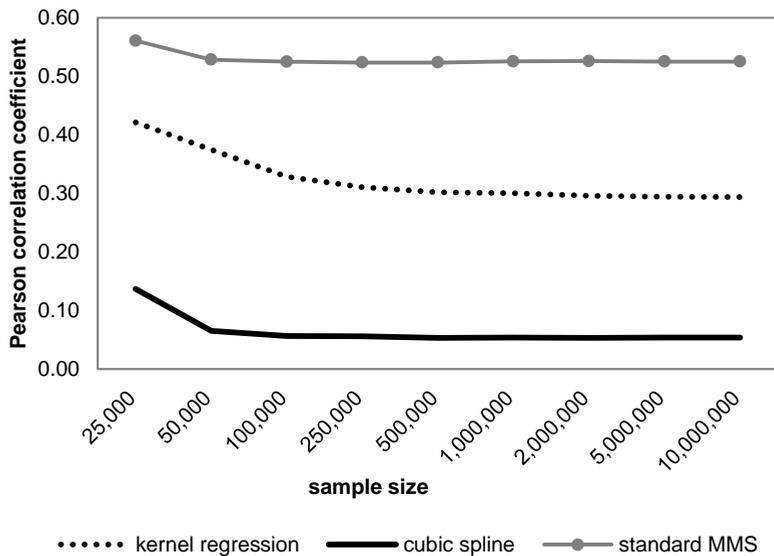
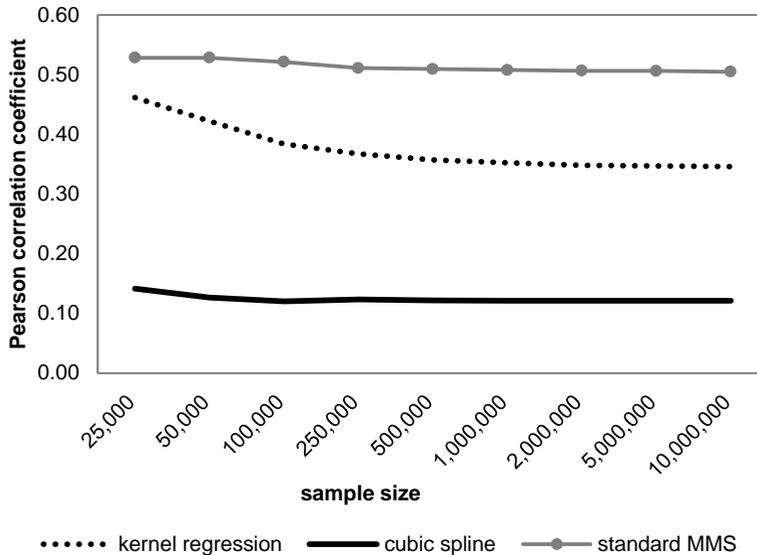


Figure 6b: First-degree autocorrelation of average residuals, inter-provincial migration

To summarise, kernel regressions generate consistently low *AvgSS*, but squared bias is slightly higher than for cubic splines. Kernel estimates are not affected by high variance for samples with fewer than 100,000 observations, whereas cubic splines are. However, kernel estimates are problematic for the high-curvature portion of the curve at young adult ages. Cubic splines, on the other hand, generate the smallest *AvgSSR* for samples with more than 500,000 observations and consistently low squared bias. While they are not affected by autocorrelation, estimates from cubic splines are affected by high variance for samples smaller than 100,000 observations, particularly at older ages. By contrast, standard MMSs perform poorly according to all goodness-of-fit indicators because they constrain migration intensity to follow a constant rate of ascent and descent and locate the estimated peak value of the age profile between the student and the labour force curves. To take into account the increase in migration intensity in the early twenties observed in large samples, in the next section we extend the standard MMS to include a student curve as described in Wilson (2010), which we denote “student MMS”.

5.2 Student MMS versus cubic spline and kernel regression

We now fit a student MMS to samples with at least 250,000 observations, which is the smallest sample size required to distinguish the presence of a student curve in Chile. Figure 7 compares average inter-provincial migration age profiles fitted with standard and student MMSs. It clearly shows that student MMSs generate an excellent fit that closely matches the true distribution, particularly along the high-curvature portion of the curve at young adult ages, and that it identifies both the student and labour peaks. Figures 8a and 8b represent *AvgSSR* by model and sample size and confirm that student MMSs provides the best fit for all available sample sizes, although for migration between municipalities it generates *AvgSSR* very similar to that of cubic splines for samples with at least 5 millions observations.

Figure 7: Inter-provincial migration age profiles, selected sample sizes

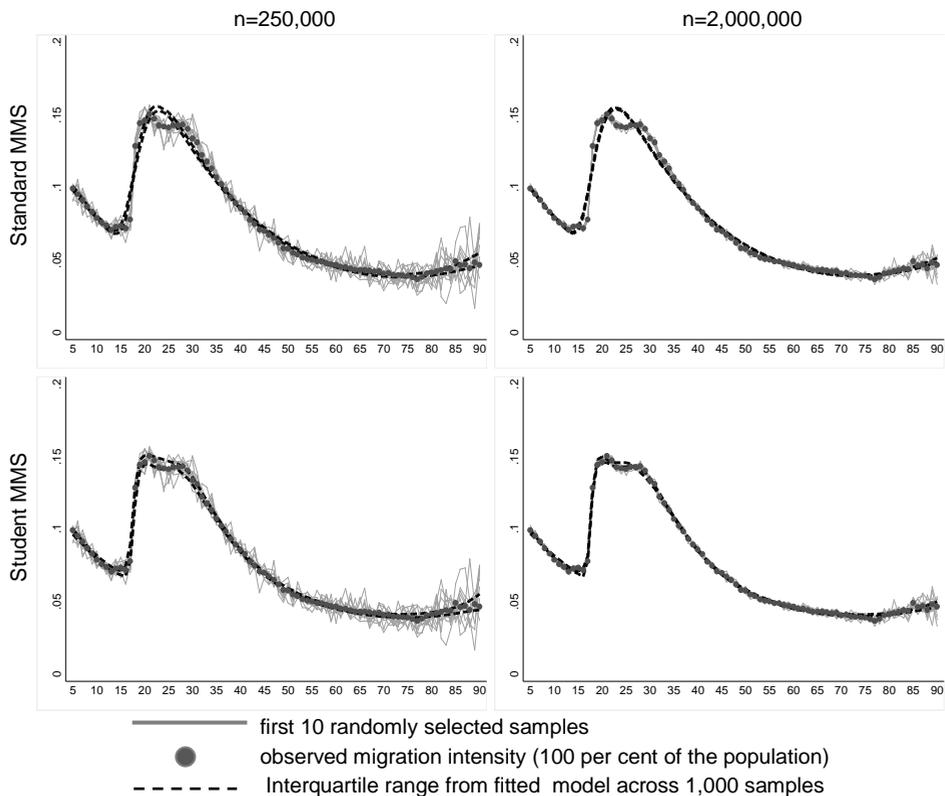


Figure 8a: Log of *AvgSSR* by sample size, migration between municipalities

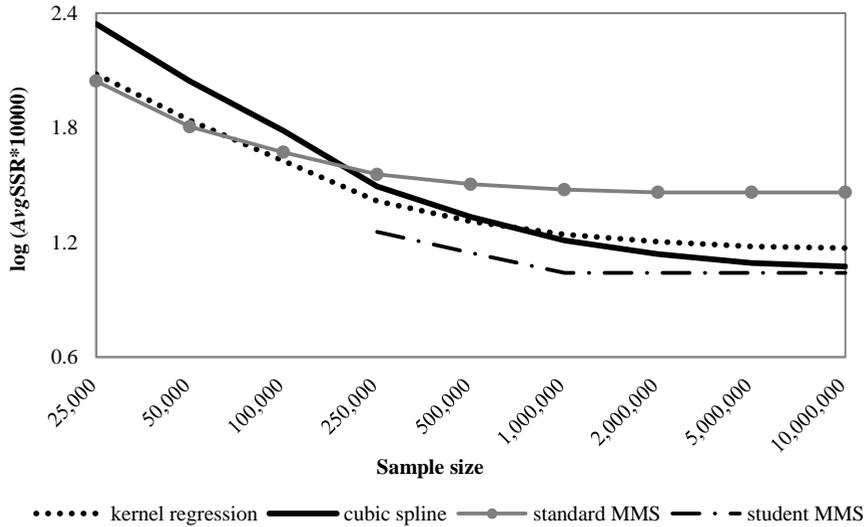
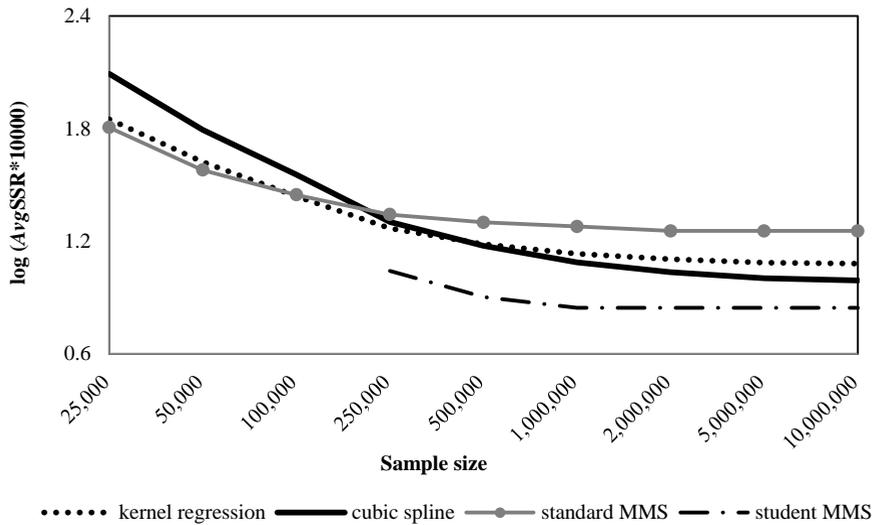


Figure 8b: Log of *AvgSSR* by sample size, migration between provinces



6. Research application

The shape of the fitted curve fundamentally affects the specific features of the migration age profile used for comparative purposes. Since the 1980s a large body of literature has developed that compares migration age patterns across countries (Bell and Muhidin 2009; Kawabe 1990; Rogers and Castro 1981) and over time (Bernard et al. 2014c; Ishikawa 2001; Rogers and Rajbhandary 1997). How well do the different smoothing techniques allow us to identify these specific features?

6.1 Comparison across countries

While model schedules have gained widespread acceptance in migration research, a number of other measures have been employed to compare migration age patterns between countries, including the migration intensity at the peak, the age at peak (Bell et al. 2002), the gross migraproduction rate (Rees et al. 2000), and the breadth of the peak (Bell and Muhidin 2009). Bernard et al. (2014a) have shown that the complexity of the migration age profile can be reduced to two principal characteristics, each of which is closely associated with other features of the age profile, and which can be summarised by two discrete indicators: the migration intensity at the peak and the age at peak. These metrics can be deduced directly from a graph or a table of smoothed age-specific migration intensities, without estimating model migration schedules.

We used estimates from Section 5 to compute the age and intensity at peak migration averaged across 1,000 Monte Carlo stimulations for each model and sample size. Tables 3a to 3d display average estimated metrics (with standard deviations in brackets) compared with the same metrics observed in the full profile. Results show that the standard MMS underestimates by two years the age at peak migration between municipalities and overestimates by two years the age peak migration between provinces. All other models provide more accurate estimates of the age at which migration between municipalities peaks. Cubic splines generate correct estimates of age at the peak for migration between provinces, whereas kernel regressions overestimate age by a year and student MMSs underestimate it by a year. We note that, at both spatial scales and for all sample sizes, the standard deviation from the average age at peak is greater for the student MMS. This indicates greater variability across estimates than for kernel regressions and cubic splines. Thus, the latter two models provide more consistent estimates of age at peak across samples. It is important to note that model migration schedules and kernel regressions allow the estimation of non-integer ages, which may be useful to discriminate countries with similar age profiles if very fine-grained estimates are required.

Turning to migration intensity at the peak, none of the models performs consistently better than the others. All models underestimate the intensity at the peak for migration between municipalities, although cubic splines generate estimates very close to the true value. For migration between provinces, student MMSs provide accurate estimates, whereas kernel regressions underestimate the peak intensity and cubic splines overestimate it. The general underestimation of metrics from kernel-based estimates can be explained by the poor fit of kernel regressions for the high-curvature portion of the curve that precedes the peak. This indicates that kernel regression should not be the preferred smoothing method to compute metrics for steep profiles. How do these results affect comparability?

Table 3a: Average age at peak, migration between municipalities

	kernel regression	cubic spline	Standard MMS	Student MMS
True value	28	28	28	28
25,000	27.9 (1.29)	27.7 (1.7)	26.69 (0.80)	-
50,000	28.0 (0.99)	28.0 (1.12)	26.24 (0.596)	-
100,000	28.0 (0.64)	28.1 (2.07)	26.18 (0.48)	-
250,000	28.0 (0.70)	28.0 (0.38)	26.1 (0.31)	27.7 (0.49)
500,000	28.0 (0.25)	28.0 (0.50)	26.0 (0.80)	27.8 (0.34)
1 million	28.0 (0.08)	28.0 (0.00)	26.0 (0.00)	27.9 (0.34)
2 million	28.0 (0.0)	28.0 (0.00)	26.0 (0.07)	27.9 (0.25)
5 million	28.0 (0.00)	28.0 (0.00)	26.0 (0.00)	28.0 (0.13)
10 million	28.0 (0.00)	28.0 (0.00)	26.0 (0.00)	28.0 (0.00)

Table 3b: Average age at peak, migration between provinces

	kernel regression	cubic spline	Standard MMS	Student MMS
True value	21	21	21	21
25,000	24.0 (2.40)	22.9 (2.69)	22.9 (1.70)	-
50,000	23.59 (2.17)	22.34 (6.08)	22.8 (0.48)	-
100,000	23.15 (1.98)	21.6 (1.61)	22.9 (0.36)	-
250,000	22.7 (1.63)	21.2 (0.97)	23.0 (0.48)	21.2 (1.92)
500,000	22.3 (1.09)	21.3 (0.97)	22.8 (0.48)	21.0 (1.76)
1 million	22.0 (0.45)	21.0 (0.40)	22.9 (0.36)	21.7 (1.54)
2 million	22.0 (0.16)	21.0 (0.00)	22.9 (0.25)	20.2 (0.98)
5 millions	22.0 (0.00)	21.0 (0.00)	23.0 (0.05)	20.5 (1.37)
10 million	22.0 (0.00)	21.0 (0.00)	23.0 (0.05)	20.2 (0.98)

Table 3c: Average intensity at peak, migration between municipalities

	kernel regression	cubic spline	Standard MMS	Student MMS
True value	0.250	0.250	0.250	0.250
25,000	0.243 (0.008)	0.255 (0.014)	0.246 (0.007)	-
50,000	0.242 (0.005)	0.250 (0.004)	0.245 (0.005)	-
100,000	0.242 (0.004)	0.249 (0.003)	0.245 (0.003)	-
250,000	0.242 (0.003)	0.249 (0.003)	0.245 (0.002)	0.247 (0.003)
500,000	0.242 (0.002)	0.249 (0.003)	0.245 (0.001)	0.247 (0.002)
1 million	0.242 (0.001)	0.249 (0.002)	0.245 (0.001)	0.246 (0.002)
2 million	0.242 (0.001)	0.249 (0.001)	0.245 (0.001)	0.246 (0.001)
5 million	0.242 (0.000)	0.249 (0.000)	0.245 (0.000)	0.246 (0.001)
10 million	0.242 (0.000)	0.249 (0.000)	0.245 (0.000)	0.246 (0.000)

Table 3d: Average intensity at peak, migration between provinces

	kernel regression	cubic spline	Standard MMS	Student MMS
True value	0.150	0.150	0.150	0.150
25,000	0.148 (0.006)	0.160 (0.132)	0.157 (0.008)	-
50,000	0.146 (0.004)	0.155 (.007)	0.155 (0.002)	-
100,000	0.146 (0.003)	0.155 (0.005)	0.155 (0.001)	-
250,000	0.145 (0.002)	0.154 (0.005)	0.155 (0.001)	0.151 (0.004)
500,000	0.145 (0.002)	0.154 (0.002)	0.155 (0.001)	0.150 (0.004)
1 million	0.145 (0.001)	0.154 (0.002)	0.155 (0.001)	0.150 (0.003)
2 million	0.145 (0.001)	0.154 (0.001)	0.155 (0.001)	0.149 (0.002)
5 million	0.145 (0.001)	0.154 (0.001)	0.155 (0.001)	0.149 (0.003)
10 million	0.145 (0.001)	0.154 (0.000)	0.155 (0.001)	0.149 (0.002)

Note: A migration intensity at peak of 0.25 indicates that 25 % of individuals aged 28 (age the peak) moved.

We illustrate the practical implication of model selection by comparing the migration age patterns of Argentina, Greece, and France. For each of the three countries we used census microdata samples with over 1 million observations from the Integrated Public Use Microdata Series (IPUMS) maintained by the Minnesota Population Center (2011). Aggregate migration age profiles were fitted to standard MMSs, cubic splines, and kernel regressions. Table 4 displays estimates of age and migration intensity at peak by model for each country. It confirms that standard MMs tend to overestimate the intensity at peak and to underestimate the age at which migration peaks by up to three years, which can significantly bias comparison across countries. Based on the age and intensity at which migration peaks in 25 nations, Bernard et al. (2014a) identified three regional migration clusters and showed that Chile exhibits age patterns very similar to those of Costa Rica, Greece, Portugal, and Spain. A two-year underestimation of the peak would have incorrectly placed Chile on a par with Canada, the United States, and France. The robustness of cross-national comparison is therefore highly dependent on the accurate estimation of summary measures. These indicators are also employed to examine the association of migration age patterns with other demographic processes. Bernard et al. (2014b) used the age and intensity at which migration peaks to

demonstrate that migration age patterns broadly mirror the age structure of the life course across 27 countries. Such applications require smoothed data that accurately capture the underlying age patterns of migration.

Table 4: Estimates of the age and migration intensity at peak

	Standard MMS		Cubic spline		Kernel regression	
	Normalised intensity at peak	Age at peak	Normalised intensity at peak	Age at peak	Normalised intensity at peak	Age at peak
Argentina	2.81	25	2.62	28	2.73	28
France	3.88	25	3.68	27	3.78	26
Greece	2.74	25	2.63	28	2.69	28

Note: The spatial scales used are as follows: departments for Argentina and France, municipalities for Greece. Migration data was normalised to sum to unity.

6.2 Comparison over time

In addition to differences between countries, evidence suggests that the age profile of migration evolves over time (Bernard et al. 2014c; Ishikawa 2001; Rogers and Rajbhandary 1997) in response to systematic changes in socio-economic conditions and in levels of national development. How well do different smoothing techniques allow us to capture these changes? To assess this, we examined the evolution of inter-provincial migration age patterns in Chile since the 1980s, based on a temporally consistent spatial framework proposed by Rowe (2013) using standard MMSs and kernel regressions⁵. Fitted age profiles were normalised to sum to unity across all ages in order to identify changes in migration age patterns independent of changes in the overall level of migration (Figures 9a and 9b). The standard MMS suggests that migration age patterns in Chile have remained fairly stable since 1982, with a slight ageing of the peak from 22 years to 23 years since 1992. While a minor decrease in intensity at the peak occurred, the overall shape appears unchanged over the three decades to 2002. Kernel regressions, on the other hand, point to substantial changes. Migration started declining at a slower rate beyond the peak at the 1992 census. The shift continued into the next decade, and the 2002 age profile revealed a flattening of the peak between ages 20 and 30, which account for nearly a third of moves. That feature is concealed by the standard MMSs, which constrain migration intensity to decline following the peak.

⁵ Estimates from cubic splines were similar to those of kernel regressions.

Figure 9a: Inter-provincial migration age profile, kernel regression

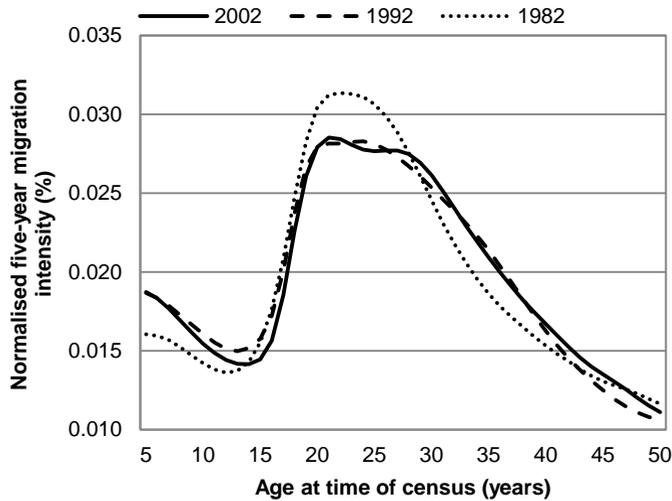
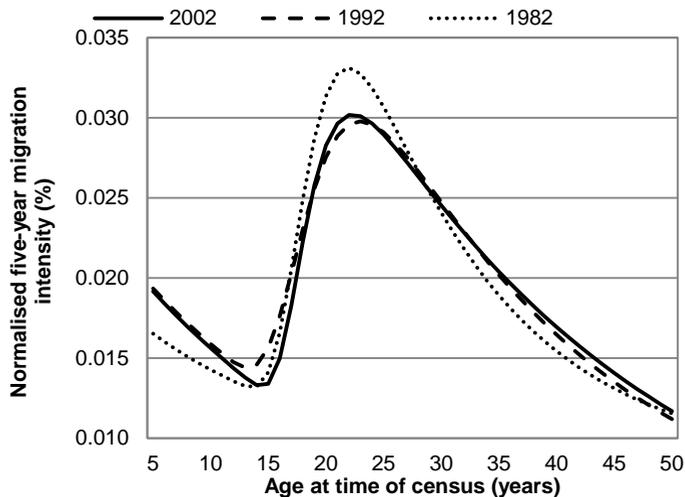


Figure 9b: Inter-provincial migration age profile, standard MMS



Source: Authors' calculations based on full population microdata from the 2002, 1992, and 1982 censuses made available by the Chilean Institute of Statistics. Migration age profiles were normalised to sum to unity across all ages and are reported from single years of age.

Model schedules have the singular advantage of being empirically based on accumulated evidence from industrialised nations and theoretically founded on the assumption that life-course transitions trigger migration at young adult ages. The unusual age patterns observed in Chile raise the question of whether kernel regression reveals a unique pattern with a clear social interpretation. The flattening of the migration peak is likely to reflect broader socio-economic changes that have occurred in recent years in Chile. Young adults typically move for education- and employment-related purposes (Mulder 1993); thus increased mobility between the ages of 20 and 30 can be explained by improved access to higher education, combined with increasing spatial concentration of education and economic opportunities in the capital city since the liberal economic reforms of the 1990s (Atienza and Aroca 2012). Tertiary enrolment doubled from 1990 to 2002, to reach 28.4% (OECD 2004). However, education opportunities have remained strongly concentrated in the capital city, where 48% of all students enrolled in higher education institutions were located in 2006 (OECD 2009). Top-tier universities are still predominantly located in the capital city, with 7 of the top 10 Chilean universities based in Santiago (Rappaport et al. 2004). Hence, young adults from regional areas are likely to move to attend university, contributing to an increase in migration intensity at young adult ages. Sustained migration intensities at high levels beyond the migration peak are probably linked to a combination of factors, including post-graduation migration and labour-related migration in early career years in a context of increasing concentration of economic activity in the capital city (Atienza and Aroca 2012). These subtle but important shifts in mobility only become apparent when a sensitive approach to the modelling of age profiles – kernel regressions or cubic splines – is employed. Other methods, in particular standard model schedules, conceal this feature by constraining migration intensity to decline following the peak.

7. Conclusion

An essential precursor to any comparative research is the establishment of valid data. In the case of internal migration, rigorous comparisons across countries and over time require a smoothing method that accurately captures the underlying age distribution of migration. The work reported here sought to evaluate the goodness-of-fit of model schedules, one of the long-standing tools of migration analysis, against two non-parametric alternatives, cubic splines and kernel regressions, and to establish the extent to which the choice of a particular method affects migration age profile indicators used for comparative research.

Conducting Monte-Carlo simulations from full population microdata for Chile, we showed that model migration schedules generate a better fit than non-parametric models when (1) the expected distribution of the age profile is known a priori, (2) the pre-determined shape of the model schedule adequately describes the true age distribution, and (3) component curves and initial parameter values can be correctly set. We showed for Chile that if any of these conditions is not met, model schedules provide a poor fit and inaccurate summary measures of the age profile. In particular, because standard model schedules constrain migration intensity to follow constant rates of ascent and descent around the peak, they tend to underestimate the age at migration, one of the key indicators used to compare migration age patterns across countries and over time (Bell et al. 2002; Bernard et al. 2014a) and to examine the association of the migration age profiles with other demographic processes (Bernard et al. 2014b). Kernel regressions and cubic splines offer more reliable alternatives by avoiding the imposition of a pre-determined distribution and by requiring little or no parameter setting.

The strength of cubic splines lies in consistently low bias and the absence of autocorrelated residuals for all sample sizes, which enables the estimation of reliable migration indicators. They are, however, affected by high variance for samples with less than 100,000 observations, resulting in higher residuals than other models. This is particularly problematic at older ages where the number of observations dwindles rapidly. On the other hand, kernel regressions are not affected by high variance, but perform poorly at estimating the high-curvature portion of the curve at young adult ages and underestimate the age and intensity at which migration peaks, which can bias comparative analysis. Caution is therefore needed in using kernel regressions to smooth migration age profiles with a sharp increase in migration intensities at young adult ages.

None of the three models examined consistently outperforms the others. The performance of a particular model depends to some extent on the specific shape of the true underlying migration age profile, which varies across countries, spatial scales, and observation periods. Unfortunately, this is rarely known a priori. Researchers therefore need to select a model according to their research aims, the characteristics of the age profile considered, and the size of available datasets.

We showed for both kernel regressions and cubic splines that the sum of squared residuals increases rapidly with decreasing sample size under 250,000 observations. However, for most countries publicly available census microdata samples are above this threshold. For example, the Integrated Public Use Microdata Series (IPUMS), maintained by the Minnesota Population Center, is the largest repository of census microdata, and at the time of writing it held 258 censuses from 79 countries. Only 6% of the available censuses contained less than 100,000 individuals, 11% had between 100,000 and 249,999 observations, 18% had between 250,000 and 499,999 observations, and the remaining 65% had at least 500,000 observations. This suggests

that kernel regressions and cubic splines can be used to reliably smooth aggregate migration age patterns of most countries. The particular significance of this finding is to enable summary indicators of the age profile, such as the age and intensity at migration peak, to be deduced directly from smoothed migration data, without estimating model migration schedules (Bernard et al. 2014a). Ease of implementation, without reliance on manual parameter-setting, makes kernel regression an important addition to the toolbox of methods for the analysis of migration and should facilitate more comparative research.

This paper has focused on aggregate migration age profiles capturing all moves within a country, regardless of distance and direction, with the aim of providing a robust and reliable basis for cross-national comparisons. Analysts are also often interested in age variations in migration streams between regions within countries. Region-to-region flows vary more widely than aggregate flows and often present problems of sample variability and small sample size. Such applications warrant greater consideration of the theoretical distribution inherent in model migration schedules, and further investigation of the applicability of kernel regressions and cubic splines to smooth inter-regional flows is desirable.

References

- Atienza, M. and Aroca, P. (2012). Concentración y crecimiento en Chile: una relación negativa ignorada. *EURE (Santiago)* 38(114): 257–277. doi:[10.4067/S0250-71612012000200010](https://doi.org/10.4067/S0250-71612012000200010).
- Beck, N., Katz, J.N., and Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42(4): 1260–1288. doi:[10.2307/2991857](https://doi.org/10.2307/2991857).
- Bell, M. and Muhidin, S. (2009). Cross-National Comparisons of Internal Migration. New York: United Nations (Human Development Research Paper 2009/30).
- Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., and Hugo, G. (2002). Cross-national comparison of internal migration: Issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165(3): 435–464. doi:[10.1111/1467-985X.t01-1-00247](https://doi.org/10.1111/1467-985X.t01-1-00247).
- Bernard, A, Bell, M., and Charles-Edwards, E. (2014a). Improved measures for the cross-national comparison of age profiles of internal migration. *Population Studies* 68(2): 179–195. doi:[10.1080/00324728.2014.890243](https://doi.org/10.1080/00324728.2014.890243).
- Bernard, A, Bell, M., and Charles-Edwards, E. (2014b). Life-Course Transitions and the Age Profile of Internal Migration. *Population and Development Review* 40(2): 231–239. doi:[10.1111/j.1728-4457.2014.00671.x](https://doi.org/10.1111/j.1728-4457.2014.00671.x).
- Bernard, A., Bell, M., and Charles-Edward, E. (2014c). *Internal Migration Age Patterns and the Life-Course: Continuity and Change, 1970–2000*. Paper present at the European Population Conference, Budapest, Hungary, 25–28 June.
- Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling* 1(3): 195–211. doi: [10.1177/1471082x0100100303](https://doi.org/10.1177/1471082x0100100303).
- Bracken, I. and Bates, J. (1983). Analysis of gross migration profiles in England and Wales: some developments in classification. *Environment and Planning A* 15(3): 343–355. doi:[10.1068/a150343](https://doi.org/10.1068/a150343).
- Carter, D.B. and Signorino, C.S. (2010). Back to the future: Modeling time dependence in binary data. *Political Analysis* 18(3): 271–292. doi:[10.1093/pan/mpq013](https://doi.org/10.1093/pan/mpq013).
- Castro, L.J. and Rogers, A. (1983). Patterns of family migration: two methodological approaches. *Environment and Planning A* 15(2): 237–254. doi:[10.1068/a150237](https://doi.org/10.1068/a150237).

- Congdon, P. (1993). Statistical graduation in local demographic analysis and projection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 156(2): 237–270. doi:10.2307/2982731.
- 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. (eds.). *International Migration in Europe: Data, Models and Estimates*. Chichester, UK: John Wiley & Sons: 193–205.
- Goodall, C. (1990). A survey of smoothing techniques. In: Fox, J. and Long, J. (eds.). *Modern methods of data analysis*. Newbury Park, CA: Sage Publications: 126–176.
- Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer-Verlag. doi:10.1007/978-3-642-17146-8.
- Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. Boca Raton, Florida: CRC Press.
- Ishikawa, Y. (2001). Migration turnarounds and schedule changes in Japan, Sweden and Canada. *Review of Urban & Regional Development Studies* 13(1): 20–33. doi:10.1111/1467-940X.00029.
- Kawabe, H. (1990). Migration Rates by Age Group and Migration Patterns: Application of Rogers Migration Schedule Model to Japan, the Republic of Korea, and Thailand. Tokyo, Japan: Institute of Developing Economies.
- Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics & Probability Letters* 7(3): 195–199. doi:10.1016/0167-7152(88)90050-8.
- McNeil, D.R., Trulless, T.J., and Turner, J.C. (1977). Spline interpolation of demographic data. *Demography* 14(2): 245–252. doi:10.2307/2060581.
- Minnesota Population Center (2011). Integrated Public Use Microdata Series, International: Versions 6.1 [Machine-readable database]. Minneapolis: University of Minnesota.
- Moguerza, J.M, Olivares, A., Kostaki, A., and Psarakis, S. (2010). *Comparing Smoothing Techniques on Fertility Data*. Paper present at the Ninth Mexican International Conference on Artificial Intelligence, 8–13 November 2010. doi:10.1109/MICAI.2010.15.
- Mulder, C.H. (1993). Migration dynamics: a life course approach [PhD thesis]. University of Amsterdam.

- OECD (2004). *Reviews of National Policies for Education, Chile*. Paris: OECD.
- OECD (2009). *OECD Territorial Reviews: Chile*. Paris: OECD.
- Pagan, A. and Ullah, A. (1999). *Nonparametric econometrics*. Cambridge: Cambridge University Press.
- Peristera, P. and Kostaki, A. (2005). An evaluation of the performance of kernel estimators for graduating mortality data. *Journal of Population Research* 22(2): 185–197. doi:10.1007/BF03031828.
- Rappaport, D., Benavent, J.M., and Meller, P. (2004). Rankings de universidades chilenas segun los ingresos de sus titulados. Banco Central de Chile. (Documento de Trabajo Numero 306).
- Rees, P., Bell, M., Duke-Williams, O., and Blake, M. (2000). Problems and solutions in the measurement of migration intensities: Australia and Britain compared. *Population Studies* 54(2): 207–222. doi:10.1080/713779082.
- 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. doi:10.1080/08898489409525372.
- 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. doi:10.1111/1467-940X.00001.
- 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. doi:10.1080/08898480590902145.
- Rogers, A., Little, J., and Raymer, J. (2010). *The indirect estimation of migration: methods for dealing with irregular, inadequate, and missing data*. Dordrecht: Springer. doi:10.1007/978-90-481-8915-1.
- Rogers, A. (1988). Age patterns of elderly migration: an international comparison. *Demography* 25(3): 355–370. doi:10.2307/2061537.
- Rogers, A. and Castro, L.J. (1981). Model migration schedules. Laxenburg, Austria: International Institute for Applied Systems Analysis (Research Report RR-81–30).

- Rogers, A. and Watkins, J. (1987). General versus elderly interstate migration and population redistribution in the United States. *Research on Aging* 9(4): 483–529. doi:10.1177/0164027587094002.
- Rogers, A. and Rajbhandary, S. (1997). Period and cohort age patterns of US migration, 1948–1993: Are American males migrating less? *Population Research and Policy Review* 16(6): 513–530. doi:10.1023/A:1005824219973.
- Rogers, A., Racquillet, R., and Castro, L.J. (1978). Model migration schedules and their applications. *Environment and Planning A* 10(5): 475–502. doi:10.1068/a100475.
- Rowe, F. (2013). Spatial Labour mobility in a transition economy: Migration and commuting in Chile [PhD thesis]. The University of Queensland
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511755453.
- Schoenberg, I.J. (1969). Cardinal interpolation and spline functions. *Journal of Approximation theory* 2(2): 167–206. doi:10.1016/0021-9045(69)90040-9.
- Sharef, E., Strawderman, R.L., Ruppert, D., Cowen, M., and Halasyamani, L. (2010). Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic Journal of Statistics* 4: 606–642. doi:10.1214/10-EJS566.
- Shryock, H. and Siegel, J. (1975). *The methods and materials of demography*. Washington, DC: US Government Printing Office.
- StataCorp (2013). *Stata Users Guide, Release 13*. College Station, TX: Stata Press.
- Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing*. London: Chapman & Hall/CRC. doi:10.1007/978-1-4899-4493-1.
- Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research* 23(8): 191–222. doi:10.4054/DemRes.2010.23.8.

