MigraR package: Performing Rogers and Castro's multi-exponential models for migration estimation

J. Sebastian Ruiz-Santacruz
ADRI, Shanghai University China, China / GRUMIS - Universidad de Narino, Colombia

Background

Migration is a phenomenon that shows strong regularities in its migratory patterns by age. The study of these patterns was introduced to read the regularities that appear in internal migration and that have been interpreted within the framework of a lifetime: for example, it is widely recognized that movements by age are dependent on each other, as the case of children who migrate with their parents at the same time, or wives with their husbands. Over time, the need arises to seek measures that offer the ability to determine with greater certainty the observations through the mathematical study of their behavior and relationship. The cause of the few studies of international migration seems to be related to the lack of data compiled by age and sex of this phenomenon.

The mathematical study of migration begins in the seventies just after having been applied in other branches of demography such as fertility, nuptiality and mortality using model patterns to describe the specific rates of these phenomena (Coale and McNeil, 1972; Coale and Trussell, 1974; Lee and Carter, 1992). Thus, Rogers, Raquillet, and Castro (1977) make the first approximations that utilize a mathematical formulation based on the shape of the curves and the relationships between their parameters taking as examples the internal migrations of the United States, Poland, and Sweden and using as reference the distance between the curve dominated by infants and the curve of the descent of the working ages, which they called parental change or parental shift.

Later, Rogers and Castro (1981) present a work in which they establish a general polynomial with 7, 11 and 13 parameters. Such seminal work focuses on the most common regularities presented by internal migration patterns by age, begins with a high concentration of young adults, followed by high migration rates among children who begin during the first years of life, falls at a low point around 15 to 20 years, increasing the rate sharply to a local maximum close to 25 to 30 years of age, and decreases regularly until the peak of retirement is shown. These regularities are expressed, in general, by a single exponential negative curve of the ages of the workforce before the birth of the children, a uni-modal curve skewed to the left of the ages of the workforce and, an almost bell shaped curve of in the ages posterior to the labor force. All these curves are parameterized as is shown in the further example for the complete model.

In this way, the study of the patterns was divided in two: the applications of the models in different populations and the improvement of the estimation of the parameters. Within the first applications, the one that simulates various scenarios is presented, where changes in the calendars are observed based on economic and demographic impulses induced on the cohort and its size (Pandit, 1997). In the same line, we can find the comparison of internal migration curves of the population born abroad in the United States (Rogers and Raymer, 1999a), or the applications on migrant flows in Europe (Rees, 1977; Raymer and Rogers, 2008). On the ways of estimation, the software is introduced to perform the estimation such as MODEL or TableCurve2D (Rogers and Raymer, 1999b), or alternatives with methods such as Nedler Mead (Nelder and Mead, 1965) implemented in Solver of the Excel software and utilized in the examples given in IUSSP (2018), or those described to find linearly the parameters of the equation (Rogers, Castro, and Lea, 2005). Further on, observations on countries in which migrants generated by the incorporation of troops into the education system and on how to describe the equation including the parameters of this curve are incorporated (Wilson, 2010). Recently Hussinger (2019), Riffe, et al. (2020) and Dyrting (2021), has tackled the problem also in the way performed (Ruiz-Santacruz and Garces, 2018) generating R code for replication and using the idea of prior distributions for initial values.

The last works that analyze the structure of the migratory calendars are those derived from the IMAGE project, more exactly those produced by Bernard, Bell, and Charles- Edwards in 2014, that although they do not intend to exclude advances on conventional estimates, utilize non parametric measures on the curves to make comparisons. In the same way, the benefits of performing different types of smoothing of the observed rates are mentioned below (Bernard and Bell, 2015), which eventually ends up being a preliminary step to make parametric estimates since it improves the optimization of parameters The most important application is consolidated when these measures are incorporated in the analysis of all the internal migrations of Latin American countries, something that tries to emulate this framework (Bernard, Rowe, Bell, Ueffing and; Charles-Edwards, 2017). Thus, the applications in which the model has been used mainly correspond to those related to the need to describe elegantly the migration rates, that is, the estimation, the graduation (or adjustment of noisy information), the comparative analysis between countries, the reduction of the volume of information to be replaced by parameters and its utilized in population projections of several countries (Bates and Bracken, 1982, 1987; Liaw and Nagnur, 1985; Mcmeekin, 1998; Potrykowska, 1988; Rogers and Raymer, 1999a; Wilson, 2010).

First of all, the influence has been analized simulating the initial parameters, taking as a reference a uniform distribution in a priori way, to get rid of the assumption of starting empirically analyzing the curves (by eye) to establish their possible initial parameters. The main objective is to find a robust solution, which does not depend on the arbitrariness of the initial values. Besides, the analysis will focus on the study of a type of migration not frequently studied as international, due to the lack of information generated among others due to the low mobility between some countries. In general, is considered that migratory calendars and the meaning they acquire in the analysis of vital events such as emacipation, studies, trawling migration, labor migration or post-labor migration, is key in academic and political analysis very present in the today's debate. Migration profiles also talk about how linked to the family are the movements of a certain system since, there is a higher propensity to migrate from family members left in the origins. This will be observed in the optimizations made for all intra-Latin American migrations. The proposed work acquires relevance to the extent that, in the first place, a simulation is presented in a simple way that releases the assumption of arbitrary and fixed initial values creating an R tool utilized to allow its replicability and second, additional results areobtained to those of simply optimizing the best model curve and checking the assumptions of the model, and the detection of a new migration pattern that can lead to a new parametrization.

The proposed work acquires relevance to the extent that, in the first place, a simulation is presented in a simple way that releases the assumption of arbitrary and fixed initial values creating an R tool utilized to allow its replicability and second, additional results are obtained to those of simply optimizing the best model curve and checking the assumptions of the model, and the detection of a new migration pattern that can lead to a new parametrization.

Why Rogers and Castro and the package?

- It estimates a migration schedule based on exponential curves that belong to a demographic behaviour based on a simulation of the initial values.
- Give us some demo-economic parameters.
- Other alternatives are kernel smothing (in case of under-5 child over estimation suspicion) and non parametrical parameters based on the curves.
- It could be adjusted to other definitions such as definitions: post-retirement migration defined by Wilson (2020)
- Some countries still need methods to smooth migration and grasp the mean behaviour of the migration calendar.
- There are opportunities to use it in estimation children under 5 years old modifiying S3 class into the package. p.e. When empirical data shows it historically.

Migration parameters

The estimated parameters have an interpretation that, under the classical model, is given in economic terms and allows a demographic analysis in which the drag that migrants of their children are commonly incorporated (take them with them when migrating), the movement in working ages and the behavior of migration in retirement ages. Thus, the most general equation for the multi-exponential model that describes the most complex parametric model of 13 parameters takes the form (1):

$$M(x) = a_1 e^{-\alpha_1 x} + a_2 e^{-\alpha_2 (x-\mu_2) - e^{(}-\lambda_2 (x-\mu 2))} + a_3 e^{-\alpha_3 (x-\mu_3) - e^{(}-\lambda_3 (x-\mu_3))} + a_2 e^{\lambda_4 x} + c$$

Where: $M(x)$ Standardized migration rate by age x, μ_2, μ_3 : Location parameters, α_1 : rate of descent of the workforce component, λ_2 : rate of rise of the workforce component, α_2 : rate of descent of the workforce component, λ_3 : rate of ascent of the post-workforce component, α_3 : rate of decline of the post-workforce component, x_l : low point, x_h : high peak, x_r : retirement peak, X : displacement of the workforce, A : parental shift, B : jump, c : constant, and, a_1, a_2, a_3, a_4 : levels and coefficients of the equation.

The optimization is carried out on the basic parameters of the model, however, Rogers and Castro (1981) propose other measures that have a more practical explanation to describe a set of migrations. Three have been chosen that would represent the interest of the knowledge of the migratory calendars that are developed within the Latin American region. On the one hand, there is the description of the existence of a dominant labor curve determined by the $\frac{a_2}{a_1}$ ratio, and its reciprocal index of child (infant) dependence, $\frac{a_1}{a_2}$. On the other hand, the asymmetry of the previous labor curve is studied, which is important to understand how the origins contribute to the working-age population, $\frac{\lambda_2}{\alpha_2}$. See Figure 1 to check the parameters. See Figure 1 to check the parameters.

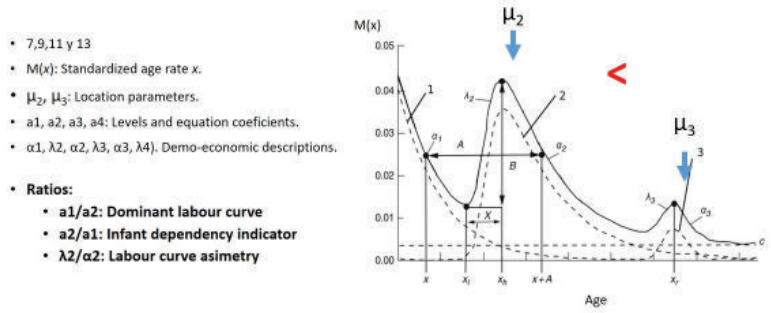

Figure 1: Based in Raymer, Rogers and Little, 2010

Procedure

One of the weak points in the estimation of migration models through optimization is the arbitrary choice of the initial parameters. In optimization processes, it is common to obtain different values of the optimal tuple of parameters when we vary the initials. So far the way to start the simulations maintains the strategy of using the same seed (in Meny cases the values studied by Rogers and Castro in 1981 for the city of Stockholm) to obtain the same result, that is, considering that the best curve should be based on the parameters of a curve reported above. For this problem, the result is tested with a goodness of fit test.

Similarly, the review of the literature exhibits Meny studies that utilize the χ^2 (chi-square) statistic as an adjustment criterion, which provides information on how good the estimated curve is employing the difference between the observed and adjusted values. The reference value to accept an estimate (called tolerance) is generally adjusted by the researcher, however, some studies have determined value of chi_2 below 0.001. The proposed initial algorithm takes as an exit criterion the difference between the estimated mean square errors between the best saved and the new estimated model, leaving by default (but susceptible to change) a difference (epsilon, e) of 0.00001 which is already very deMending. This criterion is very deMending according to the tests performed because it can be seen that, to achieve a better model, more than 40,000 simulations are needed. As long as this difference does not exist, the algorithm will not stop until a maximum established the number of iterations, which by default stops at 1000. Both e and the maximum number of iterations are likely to be changed by the user of the R package.

The initial information for each of the parameters is provided from a uniform distribution between 0 and 1 (or a priori non-informative distribution in Bayesian statistics) because we assume a total ignorance of the parameters with which the simulation should start. This simple step allows you to generate the desired number of curves until you reach the best model of the n iterations. Likewise, an "evolutionary" strategy is assumed, carrying out groups of n values and obtaining the best, generating a re-sampling that is obtained from a uniform distribution per parameter, which would make the initial parameters independent and identically distributed variables. The adjustment is test with the R square and the Mean Absolute Percentage Error (MAPE).

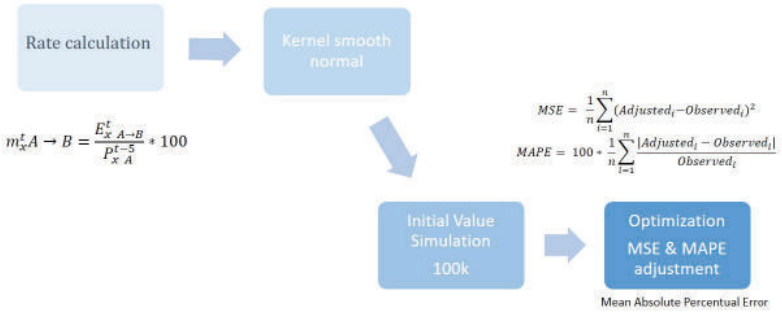

Figure 2: Estimation based in the simulation of initial values

migraR package

A least-squares optimization as a first approximation, since our objective is not to evaluate the optimization method but to emphasize two things: the simulation that begins with a priori non informative distributions using a Bayesian thought, and the comparison and grouping of the different parameters at a regional international scale. The least-squares estimates have also been used to estimate the parameters of the function composed of several exponential functions (Rogers and Little, 1994). In this way, apart from calculating the value, a confidence interval of the optimizations performed is calculated. Previously, the installation instructions of the migraR package (Ruiz-Santacruz and Garces, 2018) that implements the multi exponential models, have been made available on the GitHub web platform to make it replicable.

In terms of optimization, Rogers and Castro's functions that would model the behavior of migration are the objective functions. For any case of this objective function, what we will face is a case of a function of several variables in which, in general, there are differential calculation methods to find the maximum and minimum values. The program applies the nlminb library of the R software, which performs a linear optimization of the log of the standardized migration rate. Linear optimization is executed without restrictions, that is, the estimated parameters are not bounded. Additionally, as it starts from the points generated by the non-informative a priori distributions to perform the optimization, the program discards those points that generate a gradient equal to zero since it is not possible to find a solution.

Package example

```
library(migraR)
library(tidyverse)
data("es_asmr")
data1 <- es_asmr[-c(1,2),c(1,6)]
colnames(data1) <- c("x","y")

# Fitting and Plotting data
fitted.val.7  <- best_migramod(dataIn = data1, maxite = 200, profile = "seven")
fitted.val.9  <- best_migramod(dataIn = data1, maxite = 200, profile = "nine")
fitted.val.11 <- best_migramod(dataIn = data1, maxite = 200, profile = "eleven")
fitted.val.13 <- best_migramod(dataIn = data1, maxite = 200, profile = "thirteen")

plot(data1, cex=0.1, xlab = 'Age',
    ylab = 'Standardized Migration Rate')
lines(data1[,1], fitted.val.7$modelClass$value(fitted.val.7$bestParam,data1), col="blue")
lines(data1[,1], fitted.val.9$modelClass$value(fitted.val.9$bestParam,data1), col="orange")
lines(data1[,1], fitted.val.11$modelClass$value(fitted.val.11$bestParam,data1), col="blue")
lines(data1[,1], fitted.val.13$modelClass$value(fitted.val.13$bestParam,data1), col="green")
legend('topright',
```

```
legend = c(paste("(7)", "MAPE:", round(as.numeric(fitted.val.7$bestMAPE),2),
                 "R²:", round(as.numeric(fitted.val.7$bestRcuad),3)),
           paste("(9)", "MAPE:", round(as.numeric(fitted.val.9$bestMAPE),2),
                 "R²:", round(as.numeric(fitted.val.9$bestRcuad),3)),
           paste("(11)", "MAPE:", round(as.numeric(fitted.val.11$bestMAPE),2),
                 "R²:", round(as.numeric(fitted.val.11$bestRcuad),3)),
           paste("(13)", "MAPE:", round(as.numeric(fitted.val.13$bestMAPE),2),
                 "R²:", round(as.numeric(fitted.val.13$bestRcuad),3))),
         col = c("red",'orange',"blue","darkgreen"), lty = c(2,6,3,5))
```

More information: https://github.com/elflacosebas/migraR

Results from Latin American migration system

Data visualization and estimation

- Data on Latin American migration: REDATAM project https://bit.ly/3qf454D
- Most of the models adjusted ended in 11 or 13 parameters (115 out of 139 pairs of origin-destination).
- Many curves do not fit with the pattern because there is a curve between 5 and 15 aprox that we have called: Child migration delay.
- Candidate to be a mechanism of migration (internal or international).
- Worst estimation of the post-retirement migration peak.

Examples

Using the log of the estimated ratios we can say that the latinamerican migration system has a medium-high child dependency, a predominance of labour force curve and a very asymetric labour force curve, are congruent with the reality of Latin American migrations. See Figure 3 and Figure 4 to see the example of the adjustments. Figure 5 shows a typology made from a statistical cluster made with the package hclust in R. Figure 6 contains the logarithm of the ratios calculated to obtain a whole picture of the latin american system.

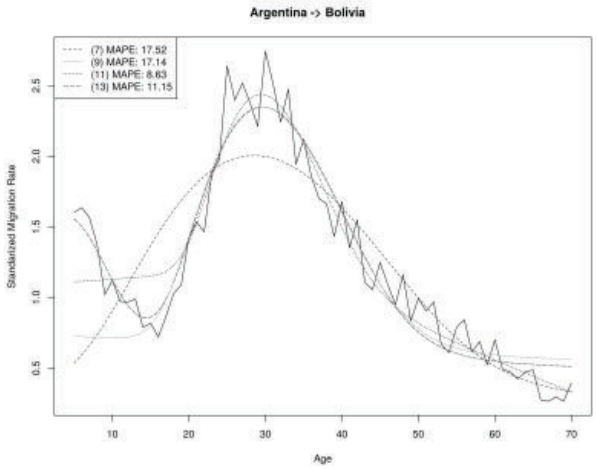

Figure 3: Curves for male and the estimated MAPE and R square, from Argentina to Bolivia

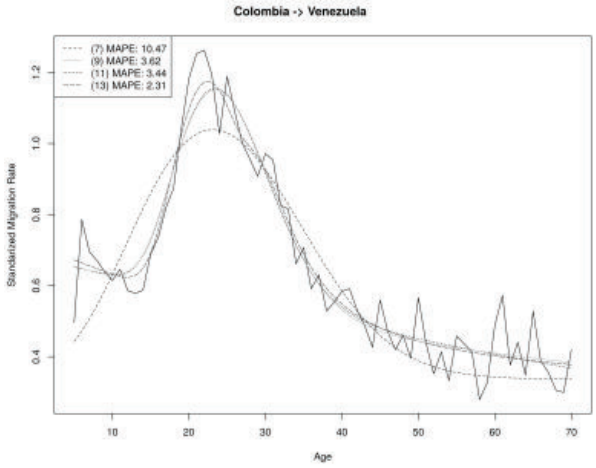

Figure 4: Curves for male and the estimated MAPE and R square, from colombia to Venezuela

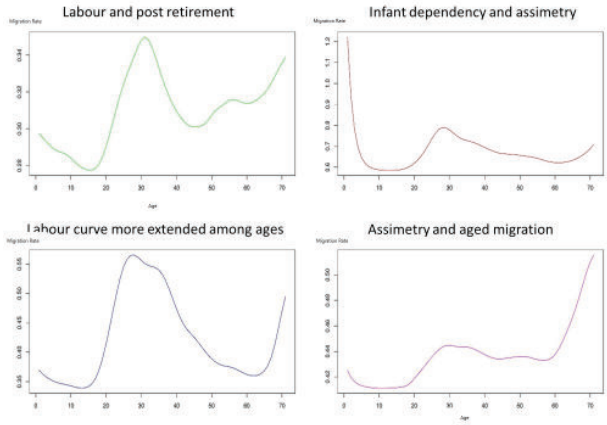

Figure 5: Typology of migration build from cluster of parameters

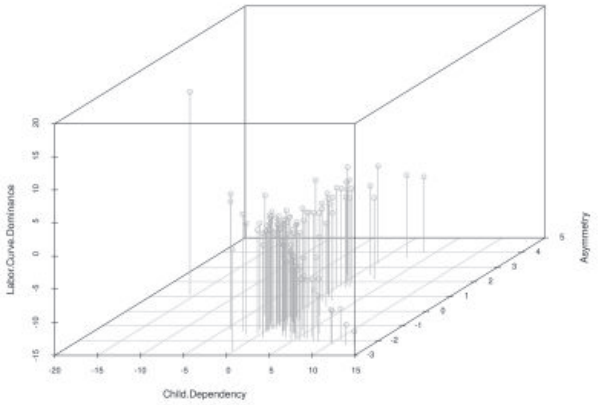

Figure 6: Log of parameters for the estimated curves

In Figure 3 we can see the way in which the fitted curves can correspond to the observed data (tested with the MAPE), however, the curve also exhibits a negative slope at ages under 10 year and a better fit of a model of 11 parameters. Meanwhile, the Figure 4 shows a local maximum before the age 10 that cannot be adjusted, and necessarily leads us to a model with 13 parameters, preventing the adjustment of a possible peak between the ages of 60 and 70 years old. Even so, the fitted models show smaller errors than required to be considered a good fit.

In relation to figure 5, we can see that the statistical grouping procedure made with the hclust package in R shows four sets of countries differentiated by their way of emigrating. All of them show a curve in the economically active ages, however those that show more intensity are the group in which an additional post-retirement curve is shown (top-left). The other group exhibits a labor curve is quite wide and extends over various ages (bottom-left). then there is another curve that shows a very strong migration intensity for children under 10 years of age that is accompanied by a less prominent labour force curve, and a greater asymmetry that extends almost to 60 years (top-right). Finally, it shows a curve in which there is asymmetry in the labor curve, but there is a fairly intense post-retirement migration (bottom-right).

On the other hand, the Figure 6 shows us the logarithm of the calculated ratios between parameters and how as child dependency increases while the dominance of the labor curve and symmetry decrease. This could give us the possibility of saying that there is may be a demographic effect, when considering that people who mostly emigrate with children can do so at older ages and with the willingness to work until later in life.

Conclusions

1. The first of the conclusions revolves around the estimation of the parameters using simulations based on a priori non-informative distributions for the parameters. In this sense, we always find an estimate, which adjusts the curve very well to the observed data. However, the location parameters μ_2 and μ_3 do not respond Meny times to the interpretations presented by the original idea of Rogers and Castro, the little signs that these parameters acquire in the determination of the Mean Square Error, is a sample of the low influence on the location of the best estimate.
2. The second conclusion to highlight is the existence of an anomaly that would skew the results in Latin American curves, which describe real bell behaviors when starting the calendar. This is a deviation from the model curves that can vary the value of the parameters when using part of the equation to describe said initial bell, falling short perhaps to describe the post-work curve so that the equation utilized is 13 parameters.
3. A simple analysis of correlations and main components viverify the mutual dependence of the parameters in the curve estimation.
4. An important contribution of this work is developed in the field of optimizations of this type of migration data observed by age because it is studied for a large number of curves, Meny coming from scarce data but allowing its modeling. In this way, we model anomalous behaviors outside of what we know as

the 'model' curve and from which still valuable information can be extracted in the joint diagnosis. However, the fact of the existence of better optimization methods using the same kind of simulation could be applied to the initial values.

5. For organisms like the International Labour Organization, it is necessary to establish robust historical measures that allow their projection and allow to carry out public policy scenarios. Besides, it could be part of a country classification according to migration characteristics that lead to better standardization. 6.Finally, here is a lack of knowledge of labor migration to group countries according to their migratory characteristics, which would give more focus to the efforts of governments in the attention and incorporation of the most vulnerable population. Knowing various estimation methodologies provides feedback for the collection and harmonization data methods.

References

1. Bates, J. and Bracken, I. (1982). Estimation of migration profiles in England and Wales. Environment and Planning A, 14(7), 889.
2. Bates, J. and Bracken, I. (1987). Migration Age Profiles for Local Authority Areas in England, 1971 to 1981. Environment and Planning A, 19(4), 521 to 535.
3. Bernard, A. and Bell, M. (2015). Smoothing internal migration age profiles for comparative research. Demographic Research, 32(1), 915 to 948.
4. Bernard, A., Bell, M., and Charles Edwards, E. (2014). Improved measures for the cross national comparison of age profiles of internal migration. Population Studies, 68(2), 179 to 195.
5. Bernard, A., Rowe, F., Bell, M., Ueffing, P., and Charles Edwards, E. (2017). Comparing internal migration across the countries of Latin America: A multidimensional approach. PLoS ONE, 12(3).
6. Coale, A. J., and McNeil, D. R. (1972). The distribution by age of the frequency of first marriage in a Women cohort. Journal of the American Statistical Association, 67(340), 743 to 749.
7. Coale, A. J., and Trussell, T. J. (1974), Model Fertility Schedules: Variations in the Age Structure of Childbearing in HuMen Populations, Population Index, 40, 185 to 258.
8. Dyrting Sigurd (2020). Smoothing migration intensiies with P-TOPALS. Demographic research. Volume 46, Article 55, Pages 1607 to 1650.
9. Fraley C. and Raftery A. E. (2002) Model based clustering, discriminant analysis and density estimation, Journal of the American Statistical Association, 97 of 458, pp. 611 to 631.
10. Fraley C., Raftery A. E., Murphy T. B. and Scrucca L. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington.
11. Hussinger, E. (2020). Applied Demography Toolbox: Eddies R Code for Fitting The Multi-Exponential Model Migration Schedule with Student Peak. Retrieved: 20 03 2020.
12. IUSSP. (2018). Tools for Demographic Estimation. Retrieved: 20 03 2020.
13. Lee, R. y Carter, L. (1992) Modeling and Forecasting U.S. Mortality. Journal of the American Statistical Association, 87 of 419, 659 to 671.
14. Liaw, K.L., and Nagnur, D. N. (1985). Characterization of metropolitan and nonmetropolitan outmigration schedules of the Canadian population system, 1971 to 1976. Canadian Studies in Population, 12(1), 81.
15. Mcmeekin, R. W. (1998). Estadisticas Educativas en America Latina y el Caribe Informe de un estudio sobre la situacion de las estadisticas educativas, indicadores y sistemas de informacion para la administracion en la region y lecciones a aprender de otras regiones. Retrieved: 20 03 2020.
16. McNeil, D. R., Trussell, T. J., and Turner, J. C. (1977). Spline Interpolation of Demographic Data. Demography, 14(2), 245.
17. Nelder, J. A., Mead, R., Nelder, B. J. A., and Meadf, R. (1965). A simplex method for function minimization. The Computer Journal, 7(4), 308 to 313.
18. Pandit, K. (1997). Cohort and Period Effects in U.S. Migration: How Demographic and Economic Cycles Influence the Migration Schedule. Annals of the Association of American Geographers, 87(3), 439 to 450.

19. Potrykowska, A. (1988). Age patterns and model migration schedules in Poland. Geographia Polonia. vol. 54, pp. 63 to 81.

20. R Development Core Team. 2012. R: A language and environment for statistical computing: Reference Index. Vienna, Austria: R Foundation for Statistical Computing.

21. Raymer, J., & Rogers, A. (2008). International Migration in Europe: Data, Models and Estimates. Applying Model Migration Schedules to Represent Age Specific Migration Flows. pp. 175 John Wiley and Sons, Ltd.

22. Riffe, T. and Aburto, J.M. and Alexander, M. and Fennell, S. and Kashnitsky, I. and Pascariu, M. and Gerland, P. (2020). Demotools: Tools for aggregate demographic analysis. Retrieved: 20 03 2020.

23. Rees, P. H. (1977). The measurement of migration, from census data and other sources. Environment and Planning A, 9(3), 247 to 272.

24. Rogers, A., and Castro, L. J. (1981). Model migration schedules. IIASA Research Report, 81, 1 to 160.b

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

26. Rogers, A., and Little, J. S. (1994). Parameterizing Age Patterns of Demographic Rates with the Multiexponential Model Schedule. Mathematical Population Studies, 4(3), 175 to 195.

27. Rogers, A., Raquillet, R., and Castro, L. J. (1977). Model Migration Schedules and Their Applications. IIASA Research Mmorandum.

28. Rogers, A., and Raymer, J. (1999a). Estimating the regional migration patterns of the foreign-born population in the United States: 1950-1990. Mathematical Population Studies, 7(3), 181 to 216.

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

30. Rogers, A., Raymer, J., & Little, J. (2010). The Indirect Estimation of Migration Smoothing Age and Spatial Patterns. pp. 47-85.

31. Ruiz Santacruz, J. and Garces, J. (2018). migraR: prototype package for R. Retrieved 20 03 2020.

32. Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, The R Journal, 8(1), pp. 205 to 233.

33. Wilson, T. (2010). Model migration schedules incorporating student migration peaks. Demographic Research, 23(8), 191 to 222.