Estimating bilateral migration flows to match known net migration totals.

Guy J. Abel & Peter W. F. Smith

9/27/2019

Abstract

Bilateral migration data, summarizing the number of people migrating between each origin and destination, provide a clearer understanding of migration patterns than summary measures such as net migration. However, bilateral migration flow data are commonly unavailable, not up to date or provide conflicting accounts of population change when compared to changes implied in demographic data.

In this paper we introduce an conditional maximisation routine to update bilateral migration data to match known net migration totals. The routine can also be applied in missing data situations to generate synthetic bilateral migration flows to match known net migration totals where no reported bilateral data are available. We illustrate the method using real world data from the United States. The resulting estimates of bilateral migration flows are demographically consistent with changes in reported population totals, births and deaths over the period and provide a detailed depiction of contemporary state to state migration patterns.

Background

One of the oldest migration estimation problems is based on estimating bilateral migration flows to match a set of known total inflows and total outflows. One solution to this problem is based on a spatial interaction model for the number of migrant in transition from origin i to destination j, during the respective time interval;

$$y_{ij} = \alpha_i \beta_j m_{ij}$$

where y_{ij} is the expected number of migrants in transition from origin i to destination j and $i, j = 1, 2, \ldots, R$ for R origins and destinations. The α_i and β_j parameters relate to the total outflow and inflow from each origin and destination respectively. The m_{ij} factor represents some auxiliary information on migration flows. This is typically additional data related to migration between the same origins and destinations. Willekens (1999) noted, in conventional spatial interaction analysis, $m_{ij} = F(d_{ij})$ where d_{ij} is a measure of distance between i and j and $F(.)$ is a distance deterrence function. Such distance deterrence functions can come in different forms, such as $F(d_{ij}) = d_{ij}^{-\epsilon}$ or $F(d_{ij}) = \exp(-\epsilon d_{ij})$, where $\epsilon > 0$ is a distance sensitivity parameter, see Sen and Smith (1995), pp. 4. Alternative specifications for m_{ij} might be travel costs or past migration flows.

As described by Willekens (1999), the estimation of parameters in a spatial interaction model can be performed by re-expressing the spatial interaction model of above in terms of a log-linear model:

$$\log y_{ij} = \log \alpha_i + \log \beta_j + \log m_{ij},$$

where unlike standard log-linear models, no intercept is included, and the final term is commonly referred to as an offset. The maximum likelihood estimates for the α_i and β_j parameters of this model can be derived using an iterative solutions to the partial differentials for the the Poisson distribution function for y_{ij}:

$$\hat{\alpha}_i = \frac{n_{i+}}{\sum_j \hat{\beta}_j m_{ij}} \quad \text{and} \quad \hat{\beta}_j = \frac{n_{+j}}{\sum_i \hat{\alpha}_i m_{ij}}.$$

Iterative solutions to these differentials are feasible when the sufficient statistics (n_{i+} and n_{+j}) are known and form an Iterative Proportional Fitting Procedure (IPFP)

Adapting Spatial Interaction Model for Known Net Migration Totals

The spatial interaction model above can be adapted to ensure that the net migration totals are the sufficient statistics of an iterative solution by having only one parameter for each region:

$$y_{ij} = \alpha_i \alpha_j^{-1} m_{ij}$$

This can be expressed as a log-linear model:

$$\log y_{ij} = \log \alpha_i - \log \alpha_j + \log m_{ij}$$

for which the maximum likelihood estimates of α can be derived by considering the probability of observing n_{ij} migrant transitions during a unit interval, given by the Poisson distribution function:

$$P(N_{ij} = n_{ij}) = \frac{y_{ij}^{n_{ij}}}{n_{ij}!} \exp(-y_{ij}).$$

The likelihood function for $\mathbf{Y} = \{y_{ij}, i, j, = 1, \ldots, R\}$ given $\mathbf{n} = \{n_{ij}, i, j, = 1, \ldots, R\}$ migrations, provided that migrations are independent, is

$$L(\mathbf{Y}; \mathbf{n}) = P(N_{11} = n_{11}, N_{12} = n_{12}, \ldots, N_{RR} = n_{RR}) = \prod_{ij} \frac{y_{ij}^{n_{ij}}}{n_{ij}!} \exp(-y_{ij})$$

Inserting model () into the likelihood function and taking the logarithmic transformation gives the log-likelihood function:

$$\begin{aligned} l(\boldsymbol{\alpha}; \mathbf{n}) &= \sum_{ij} \left\{ n_{ij} \log(\alpha_i \alpha_j^{-1} m_{ij}) - \alpha_i \alpha_j^{-1} m_{ij} - \log(n_{ij}!) \right\} \\ &= \sum_i n_{i+} \log(\alpha_i) - \sum_j n_{+j} \log(\alpha_j) - \sum_{ij} \alpha_i \alpha_j^{-1} m_{ij} + c, \end{aligned}$$

where $\boldsymbol{\alpha} = \{\alpha_h, h = 1, \ldots, R\}$, $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$ are the marginal totals, and

$$c = \sum_{ij} n_{ij} \log(m_{ij}) - \sum_{ij} \log(n_{ij}!).$$

In order to obtain the estimates of α that maximise this likelihood, we can expand the summations in ()

$$\begin{aligned} l(\boldsymbol{\alpha}; \mathbf{n}) =& n_{1+} \log(\alpha_1) + n_{2+} \log(\alpha_2) + \ldots + n_{R+} \log(\alpha_R) - \\ & n_{+1} \log(\alpha_1) - n_{+2} \log(\alpha_2) - \ldots - n_{+R} \log(\alpha_R) + \\ & (\quad \alpha_1 \alpha_1^{-1} m_{11} \quad + \quad \alpha_1 \alpha_2^{-1} m_{12} \quad + \quad \ldots \quad + \quad \alpha_1 \alpha_R^{-1} m_{1R} \quad + \\ & \quad \alpha_2 \alpha_1^{-1} m_{21} \quad + \quad \alpha_2 \alpha_2^{-1} m_{22} \quad + \quad \ldots \quad + \quad \alpha_2 \alpha_R^{-1} m_{2R} \quad + \\ & \quad \ldots + \\ & \quad \alpha_R \alpha_1^{-1} m_{R1} \quad + \quad \alpha_R \alpha_2^{-1} m_{R2} \quad + \quad \ldots \quad + \quad \alpha_R \alpha_R^{-1} m_{RR} \quad) \quad + \\ & c. \end{aligned}$$

We may then obtain the likelihood equations by differentiating with respect to a given α_h:

$$\begin{aligned}\frac{\partial l}{\partial \alpha_h} =& n_{h+}\alpha_h^{-1} - n_{+h}\alpha_h^{-1} + \\ & \alpha_1^{-1}m_{h1} + \alpha_2^{-1}m_{h2} + \ldots + \alpha_R^{-1}m_{1R} + \\ & \alpha_1\alpha_h^{-2}m_{h1} + \alpha_2\alpha_h^{-2}m_{h2} + \ldots + \alpha_R\alpha_h^{-2}m_{1R},\end{aligned}$$

where terms based on the differential for $\alpha_h\alpha_h^{-1}m_{hh}$ in () are dropped. Setting the likelihood equation to zero, multiplying through by α_h^2, and rearranging;

$$\begin{aligned}\alpha_h^{-1}(n_{h+} - n_{+h}) + \sum_{g\neq h}\alpha_g^{-1}m_{hg} + \alpha_h^{-2}\sum_{g\neq h}\alpha_g m_{gh} = 0 \\ \alpha_h^2\sum_{g\neq h}\alpha_g^{-1}m_{hg} - \alpha_h(n_{+h} - n_{h+}) - \sum_{g\neq h}\alpha_g m_{gh} = 0,\end{aligned}$$

allows us to obtain maximum likelihood estimators of α_h using the quadratic equation:

$$\hat{\alpha_h} = \frac{(n_{+h} - n_{h+}) \pm \sqrt{-(n_{+h} - n_{h+})^2 + 4\sum\limits_{g\neq h}\alpha_g^{-1}m_{hg}\sum\limits_{g\neq h}\alpha_g m_{gh}}}{\sum\limits_{g\neq h}\alpha_g^{-1}m_{hg}},$$

where α_h must be the positive root to ensure non-negative migration flow estimates. Note, the sufficient statistics (n_{h+}, n_{+h}) are unknown in our primary data. However, the difference between these two, $n_{+h} - n_{h+}$, is the known net migration total for region h and thus the corresponding parameter can be estimated. Since equation () is not a closed form expression, a direct estimate of α_h cannot be obtained. Instead, an iterative solution can be used where all initial estimates of $\hat{\alpha}_h^{(0)}$ are set to an arbitrary starting value such as unity. These are then used to update

$$\hat{\alpha_h}^{(t+1)} = \frac{(n_{+h} - n_{h+}) \pm \sqrt{-(n_{+h} - n_{h+})^2 + 4\sum\limits_{g\neq h}\alpha_g^{-1^{(t)}}m_{hg}\sum\limits_{g\neq h}\alpha_g^{(t)}m_{gh}}}{\sum\limits_{g\neq h}\alpha_g^{-1^{(t)}}m_{hg}}.$$

This is a conditional maximization of the likelihood function and converges to give estimates of α_h. Maximum likelihood estimates of y_{ij} can be obtained using model (). The iterative procedure to estimate $\hat{\alpha_h}$ and y_{ij} can undertaken using the `cm_net` routine in the *migest* R package Abel (2013).

Applied Example

In the United States, annual state to state origin-destination flow data are available from the American Community Survey (ACS), Current Population Survey or Inland Revenue Service. Each suffer from various sources of bias and inconsistencies. In addition, the US Census Bureau publishes annual domestic state net migration estimates calculated via demographic accounting.

In Figure 1 the domestic state net migration calculated from the ACS and domestic net migration reported US Census Bureau during 2017 are compared. For a few states there are large discrepancies between the reported flows. Those in the off-diagonal have differing direction of net migration implied by the net migration in each data source.

We apply our method to estimate origin-destination flows that match the US Census Bureau reported domestic net migration, using ACS state to state data as the offset (m_{ij}). The resulting flows are shown in the chord diagram of Figure 2

The differences between the estimates shown in Figure 2 and those from the ACS are relatively minor, as shown in Figure 3. However, the net migration totals of the estimates match those of the US Census Bureau.

Figure 1: Scatter plot of domestic state net migration measured by the US Census Bureau and ACS based on modulus during 2017. Axis based on log-modulus transformation.

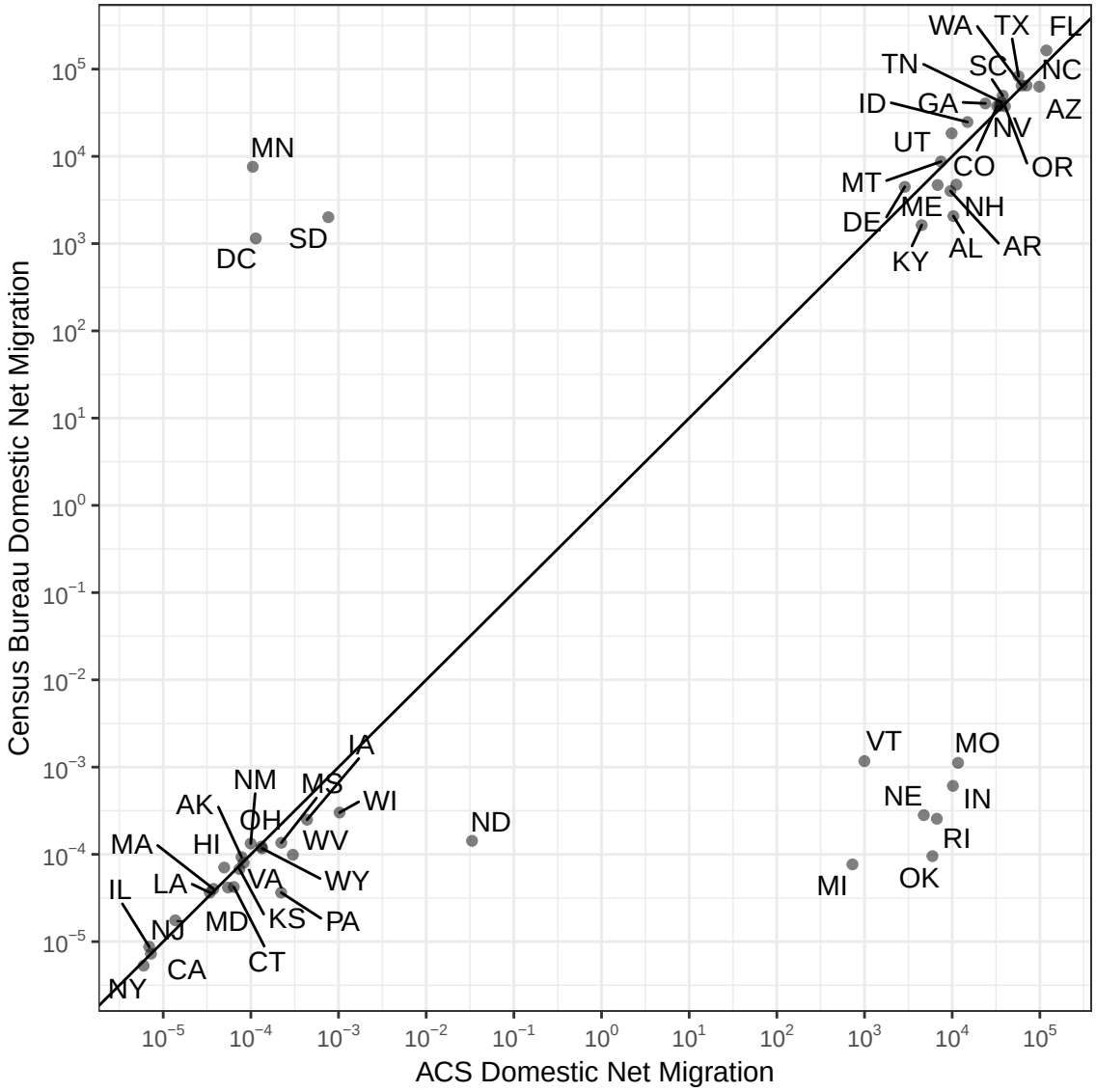


Figure 2: Chord diagram of estimated state to state migration during 2017 with net migration totals matching those reported by the US Census Bureau. Axes based on units of 100,000 persons migrating.

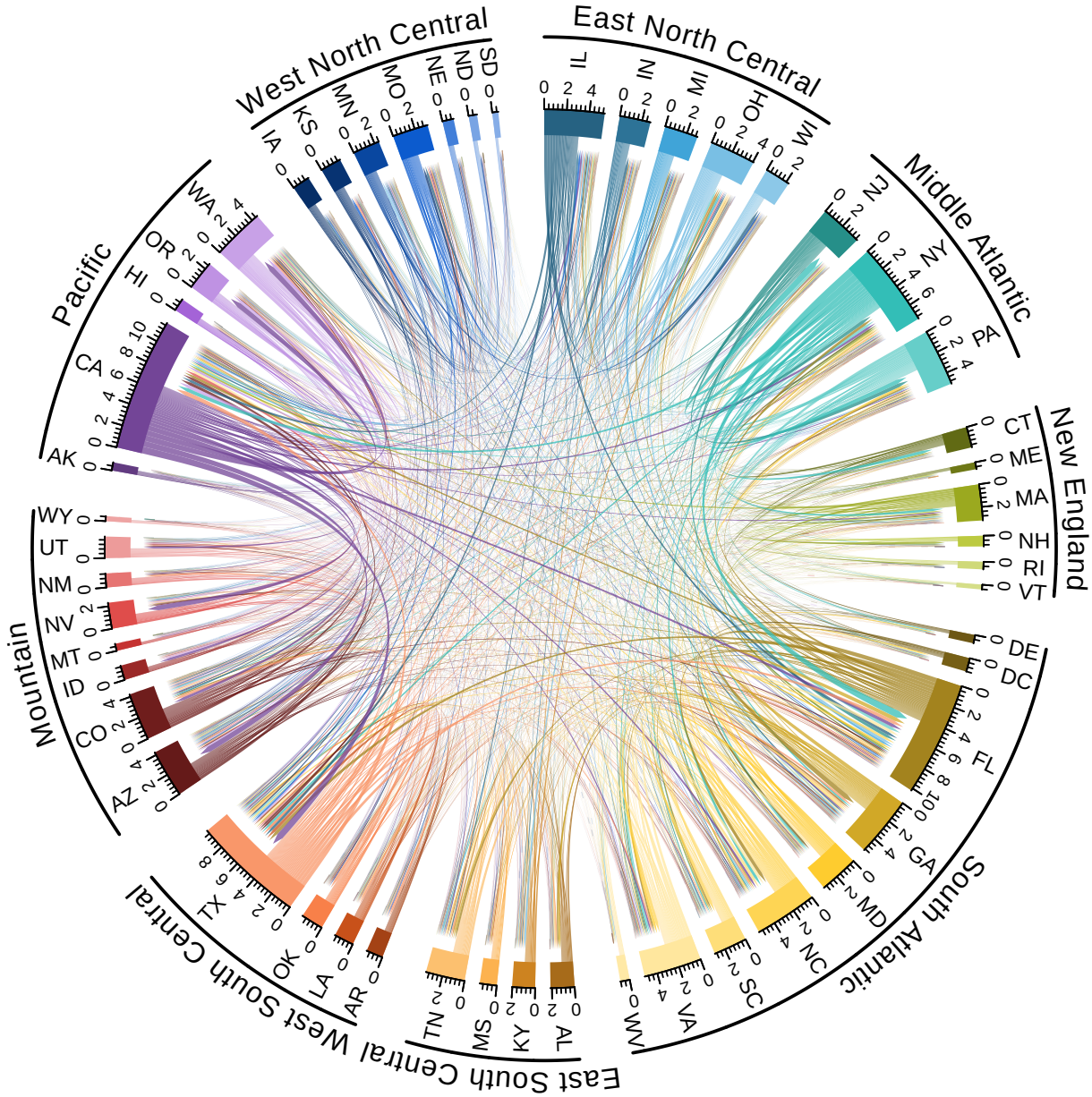
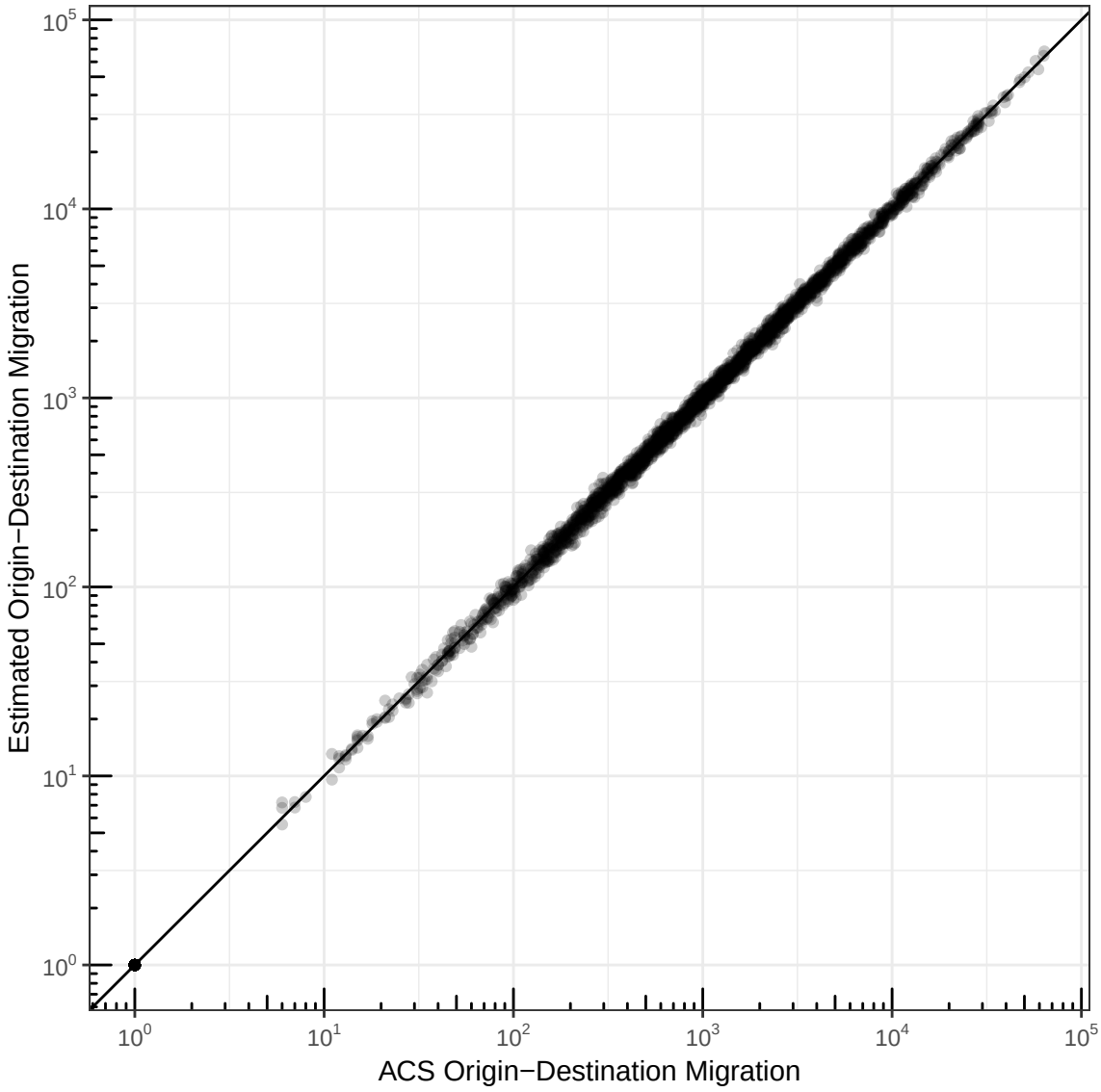


Figure 3: Comparison of estimated migration flows and ACS migration flows in 2017.



Summary

As Rees and Willekens (1986) noted, there are a number of benefits in developing population statistics that follow a demographic account. First, accounts have served economists well in their national economic modelling activities. Second, accounts force the analyst to attempt the matching of available data and a conceptual model. Third, a diverse sets of statistics are brought together in accounts and are subject to comparison and to checking for consistency. However, the same authors noted that demographic accounting has had little impact in either national statistical offices or at local or regional levels, where the usual objection posed is that the preparation of accounts tables is too complex and time-consuming an exercise. In many developed (and developing countries) these issues still persist.

In this paper we have developed a method to estimate bilateral migration flows that match known net migration totals. It provides estimated origin-destination migration flows that can form part of a full demographic account, overcoming some of the likely difficulties in producing consistent detailed migration data in countries without population registers.

References

Abel, Guy J. 2013. "migest: Methods for the Indirect Estimation of Bilateral Migration." http://cran.r-project.org/web/packages/migest/.

Rees, Philip, and Frans Willekens. 1986. "Data and Accounts." In *Migration and Settlement: A Multiregional Comparative Study*, edited by F Willekens and A Rogers, 19–58. Dordrecht, Netherlands: D. Reidel Publishing Company.

Sen, Ashish K, and Tony E Smith. 1995. *Gravity Models of Spatial Interaction Behavior (Advances in Spatial and Network Economics)*. New York, USA: Hardcover; Springer-Verlag. http://www.worldcat.org/isbn/0387600264.

Willekens, Frans. 1999. "Modeling approaches to the indirect estimation of migration flows: from entropy to EM." *Mathematical Population Studies* 7 (3): 239–78, 308. https://doi.org/10.1080/08898489909525459.