Estimation of emigration, return migration, and transit migration between all pairs of countries: Supporting Information

Jonathan J. Azose and Adrian E. Raftery

Other Supplementary Material for this manuscript includes the following:

Database S1 as separate Excel file.

Each sheet in the database contains estimates of five-year migration flows between all pairs of countries in numbers of individuals. Rows indicate country of origin and columns indicate country of destination.

1 Regional flow estimates for each five-year period

Figures 1 through 5 give regionally aggregated flow estimates for five year periods from 1990–1995 through 2010–2015. Full estimates for all pairs of countries are given in Database S1.

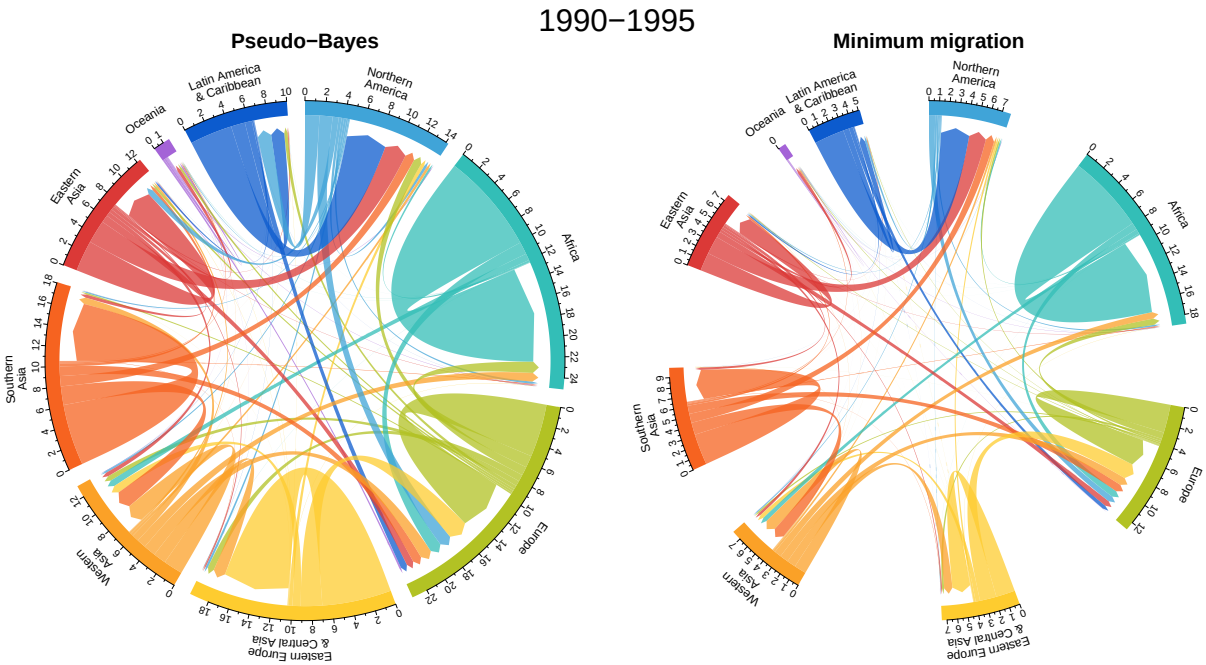

Figure 1: Estimated migration flows for 1990–1995. Left panel: Pseudo-Bayes estimates. Right panel: Minimum Migration estimates. Plots are scaled so that equal angles along the circumference of the circle represent equal numbers of migrants.

www.pnas.org/cgi/doi/10.1073/pnas.1722334115

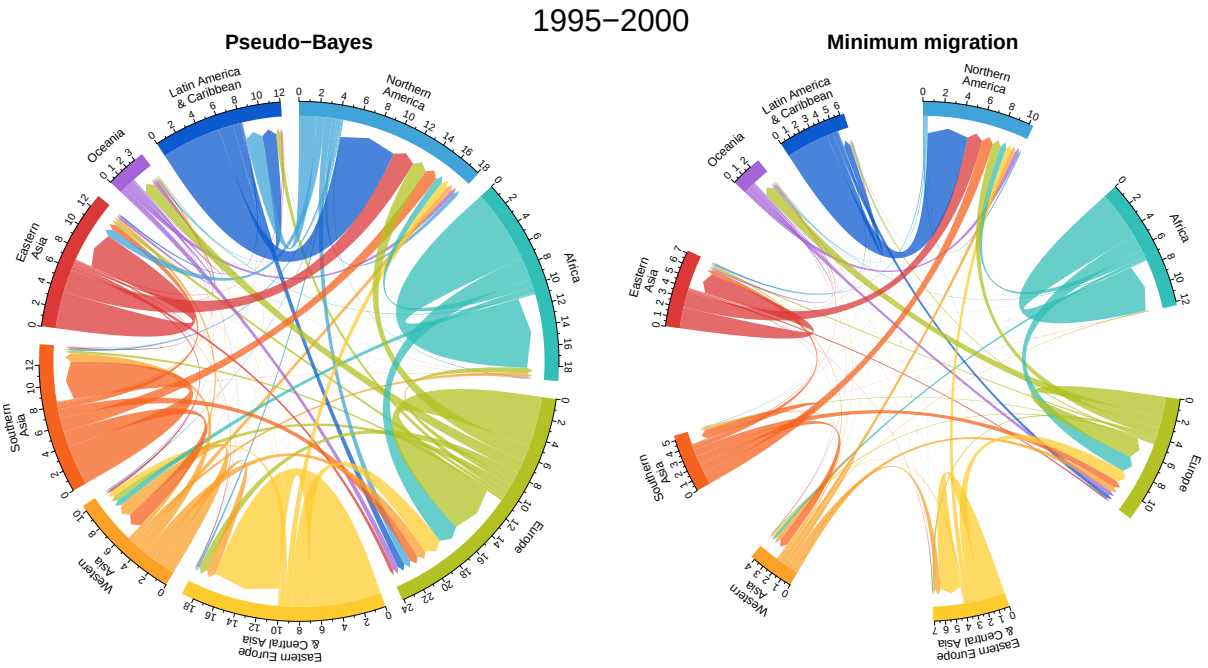

Figure 2: Estimated migration flows for 1995–2000. Left panel: Pseudo-Bayes estimates. Right panel: Minimum Migration estimates. Plots are scaled so that equal angles along the circumference of the circle represent equal numbers of migrants.

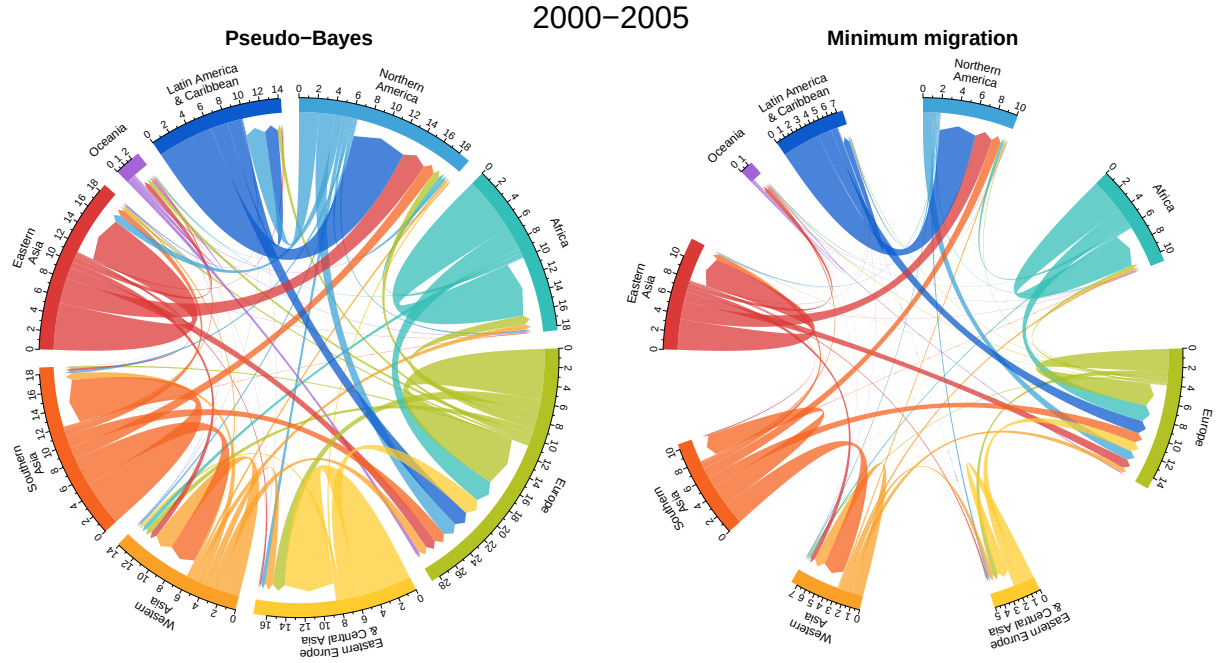

Figure 3: Estimated migration flows for 2000–2005. Left panel: Pseudo-Bayes estimates. Right panel: Minimum Migration estimates. Plots are scaled so that equal angles along the circumference of the circle represent equal numbers of migrants.

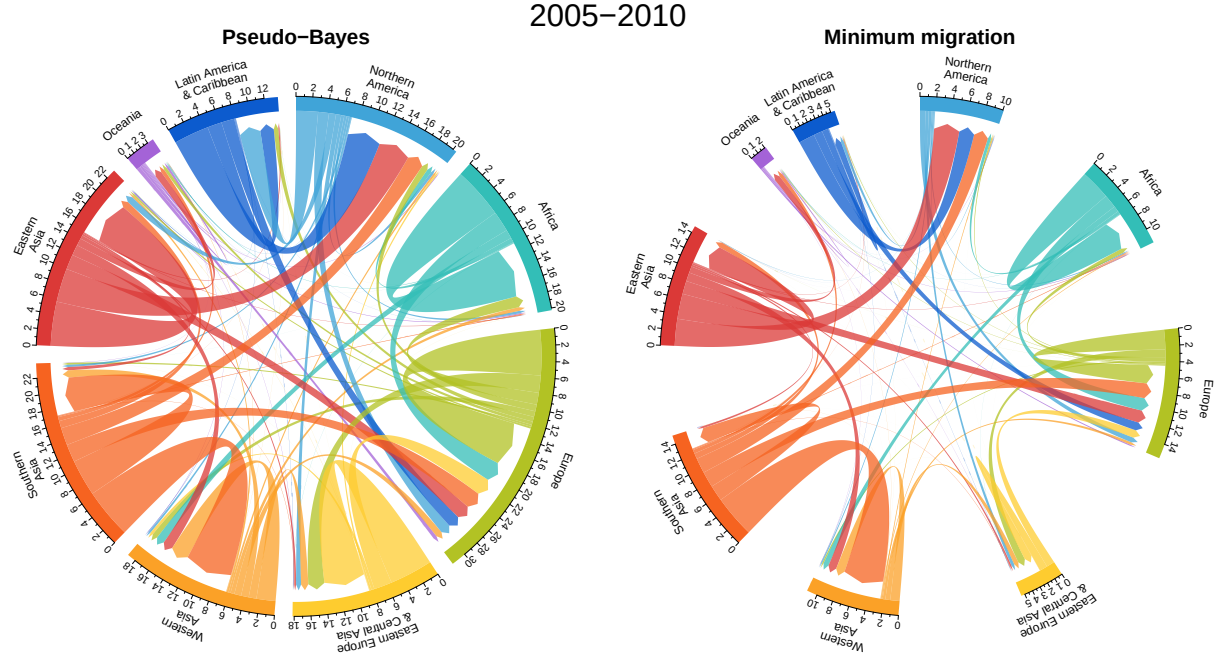

Figure 4: Estimated migration flows for 2005–2010. Left panel: Pseudo-Bayes estimates. Right panel: Minimum Migration estimates. Plots are scaled so that equal angles along the circumference of the circle represent equal numbers of migrants.

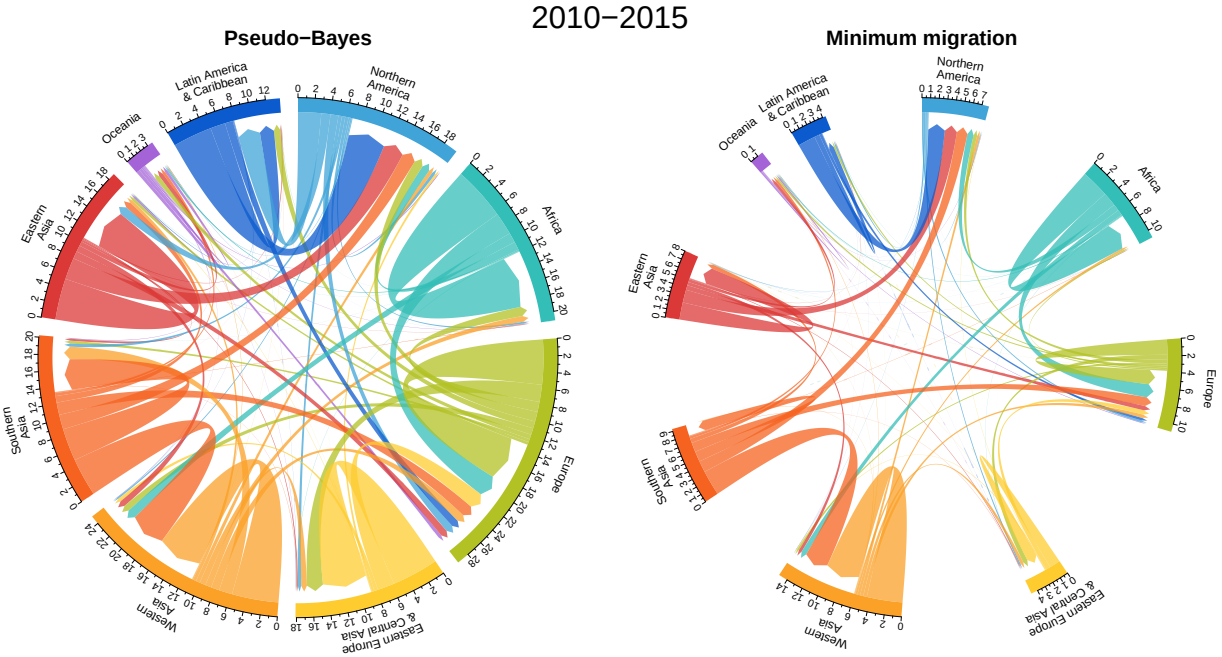

Figure 5: Estimated migration flows for 2010–2015. Left panel: Pseudo-Bayes estimates. Right panel: Minimum Migration estimates. Plots are scaled so that equal angles along the circumference of the circle represent equal numbers of migrants.

2 Additional details on pseudo-Bayes estimates

2.1 Offsets

The core Poisson model used to produce the MM estimates includes offset terms o_{ij}. In producing estimates on real data, Abel [1] sets these to

$$o_{ij} = d_{ij}^{-1}, \tag{1}$$

where d_{ij} is the distance between the capital cities of countries i and j. This offset term reflects the belief that migrant flows are more likely over shorter geographical distances, and corresponds to a gravity model originally introduced by Zipf in the context of intercity travel [2]. In the context of international migration, this belief is backed up by both theory [3, 4] and empirical evidence [5].

In the analysis in this paper, we choose instead to omit the offset terms, or equivalently, to take $o_{ij} = 1$ for all country pairs. Sensitivity analysis reveals that flow estimates produced using the MM method are extremely similar regardless of whether we choose $o_{ij} = 1$ or $o_{ij} = d_{ij}^{-1}$. We find that the correlation between estimated migration flows with the two different offsets is at least 0.999 for all time periods. This is in line with the finding by Abel [6] that including inverse distance as an offset has little impact on estimates. Given that the more complex model returns nearly identical estimates to a simpler model without offsets, we prefer the simpler model.

2.2 Estimating an optimal weight, w

Any choice of w between 0 and 1 will correspond to a valid pseudo-Bayes estimator. In practice, a sensible procedure is to select a value of w which minimizes the expected value of some loss function. Computation of a loss function requires us to have ground truth data to compare estimates against—for this application, we need a set of "true" migration flows which have been estimated via some other method.

To select w, we compare pseudo-Bayes flow estimates $M_{kt}^{PB}(w)$ against ground truth flow data from two sources: the OECD [7] and IMEM [8].

The OECD migration flow data are split into estimates of flows to and from OECD countries. Inflow data take the form of counts of the total number of foreign in-migrants to country j from all origins $\ell \neq j$ broken down by nationality k. (That is, the available quantities have the form $\sum_{\ell|\ell\neq j} m_{\ell jkt}$.) These data are on an annual time scale covering years from 2000 to 2013, although not all OECD countries have flow estimates for all years. Likewise, outflow data are estimates of the total number of foreign out-migrants from country i by nationality (i.e. $\sum_{\ell|\ell\neq i} m_{i\ell kt}$). Indices i and j are restricted to the 34 OECD countries, while nationality k covers all countries.

Note that the index k in the OECD data represents nationality rather than place of birth. For a large majority of individuals, nationality and place of birth will be in alignment. However, this will not always be the case, as individuals may change their nationality or may not receive citizenship in their country of birth. Rather than attempting to model nationality transitions on top of migratory transitions, we simply treat nationality and place of birth as interchangeable in this analysis. Flows by nationality and flows by place of birth

should be reasonably similar so long as nationality transitions and birth outside the country of citizenship are relatively rare events. The validity of the interchangeability assumption will likely vary by country, with greater validity in countries where citizenship is conferred by place of birth, or *jus soli* (as is common in the Americas), and less where citizenship is conferred by parentage, or *jus sanguinis* (as is common in most of the Eastern Hemisphere).

In contrast to the OECD data, IMEM flow estimates are broken down by country of origin and country of destination, but are aggregated over nationality or place of birth. IMEM reports aggregated flow estimates $\sum_\ell m_{ij\ell t}$ for origin-destination country pairs (i, j) where both i and j come from a set of 31 European countries. Flow estimates are provided on an annual time scale covering 2002–2008.

2.3 Selecting a loss function

A loss function, $L(m, \hat{m})$, is a function that describes the cost associated with producing an estimate of $\hat{m}$ when the true migration flow is m. In principle, the choice of loss function should be motivated by the relative costs of making different errors in estimating migration flows. As such, we will only consider loss functions with the following two properties:

1. Correctly estimating a migration flow incurs no loss. That is,

$$L(m, m) = 0, \tag{2}$$

and

2. Any estimate other than the true flow incurs more loss than the truth. That is,

$$L(m, \hat{m}) > 0 \text{ whenever } \hat{m} \neq m. \tag{3}$$

A common, convenient choice of loss function is squared error loss in migration flows, defined as

$$L_{SE}(m, \hat{m}) = (m - \hat{m})^2. \tag{4}$$

Squared error loss in flows is likely to be useful only in limited circumstances. This loss function makes the statement that the cost associated with estimating a migration flow of 10,000 when the true flow is 20,000 is the same as the cost of estimating a migration flow of 990,000 when the true flow is 1,000,000. Squared error in flows is not reasonable as a choice for a global loss function, since it will favor results that are very accurate for a few of the largest flows at the expense of misestimating many smaller flows.

It is also tempting to consider squared error in log-flows, defined as

$$L_{SEL}(m, \hat{m}) = (\log(m + k) - \log(\hat{m} + k))^2, \tag{5}$$

where k is a small, positive constant. This loss function does not suffer the same weakness as squared error loss in migration flows; the cost of estimating a flow that is 50% too small is nearly equal whether the true flow is 1,000,000 or 10,000. However, losses can get unreasonably large as true flows move towards zero. For example, if we pick $k = 1$, then $L_{SEL}(10, 2)$ is more than three times as large as $L_{SEL}(2 \text{ million}, 1 \text{ million})$. For most

practical purposes, the latter error is likely to be far more costly than the former. When dealing with very small flows, an ideal loss function would keep losses low for even large relative errors so long as absolute errors remain small.

Evaluating errors in flows on a rate scale generally allows us to avoid the weaknesses of squared error loss in flows or log-flows. We define a loss function of squared error in rates by

$$L_{SER}(m_{ij}, \hat{m}_{ij}) = \left(\frac{m_{ij}}{\text{population}_i} - \frac{\hat{m}_{ij}}{\text{population}_i} \right)^2, \tag{6}$$

where population_i denotes the total population of the origin country at the beginning of the time period. This loss function implies that there is a high cost to large errors in flow estimates relative to the population in the country of origin. Errors in small flows are not penalized heavily, as they would be if we used squared error loss on log-flows. Errors which are large relative to the true flow do not necessarily incur high loss; they are weighted heavily only when they are large relative to a country's population. All optimization in this paper minimizes over squared error in migration rates, L_{SER}.

Because of the structure of the data, it is not always feasible to put the population of the origin country in the denominator inside the loss function. For inflows to OECD countries, we use instead the population of the destination country as the denominator. For IMEM flows, where both origin and destination populations are available, we explored options using either the origin or destination population in the denominator.

2.4 Estimating a five-year weight from annual data

The optimal value of w will depend on the length of the time period in consideration. In the limit as time period length decreases to zero, we would expect the MM estimates to be perfect. (If the time period is so short that only a single person migrates, that move will be reflected by a change of one in migrant stocks, and the MM method will reproduce the single migration event.) In that limiting case, the optimal choice is to take $w = 1$ and use the correct, unsmoothed MM estimator. As the time period in consideration increases in length, we expect to see more bi-directional flows within country pairs, and therefore smaller optimal values of w.

We now derive an optimal five-year value of w based on annual data. Our derivation relies on four key assumptions—two about population change over time, and two about the distribution of migration flows within each time period. The two assumptions about population change are:

- That births and deaths have been correctly accounted for so that all population change is due to migration, and
- That neither country populations nor the composition of migrant stocks will change too dramatically over the longer time period.

The two assumptions about the distribution of transitions within each time period are both associated with the so-called "one year/five year" problem [9, 10], namely that an individual who makes multiple moves during a five-year period will be counted no more than one time

(and possibly zero times) in the five-year migration counts. Because of this effect, the true five-year transition counts must be no larger than the sum of the one-year transition counts We therefore make two fairly strong assumptions:

- That the flows within each five-year time period occur uniformly over time, and
- That the five-year transition counts are equal to the sum of the one-year transition counts. A sufficient condition to satisfy this assumption is that no individual moves more than once during the five-year period. (We will later provide details on a potential relaxation of this second assumption in Section 2.4.1.)

In reality, we do not expect these assumptions to hold precisely, but they should be a good approximation of the truth for most migration flows over a five-year period.

The quantity to be estimated is an array of five-year migration flow rates, which we will denote by M_{1-5} in order to distinguish it from annual migration flow rates $M_1, M_2, \ldots, M_5$. Finding an optimal five-year value of w is equivalent to finding a linear combination of the MM estimates $\hat{M}^A_{1-5}$ and independence estimates $\hat{M}^I_{1-5}$ which is optimal in that it minimizes mean squared error in rates. Thus our estimate is a solution of

$$\text{minimize}_{w\in[0,1]} \ \|M_{1-5} - (w\hat{M}^A_{1-5} + (1-w)\hat{M}^I_{1-5})\|^2_F, \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. For a real-valued matrix X with entries x_{ij}, the Frobenius norm of X is given by

$$\|X\|_F := \sqrt{\sum_{i,j} x^2_{ij}}, \tag{8}$$

and is equivalent to the ℓ_2-norm of the vectorized matrix.

Under the above assumptions, we can express all of the five-year quantities in terms of one-year quantities. Since flows are assumed to occur uniformly across time and with the five-year count equal to the sum of the one-year counts, we have

$$M_1 \approx M_2 \approx \cdots \approx M_5 \approx \frac{1}{5} M_{1-5}. \tag{9}$$

The annual flow counts are identical because of the assumption of uniformity across time, and must equal $\frac{1}{5}M_{1-5}$ by the second assumption. Consequently, the annual flow *rates* are approximately equal because of the assumption that country population is nearly constant over time.

How do the estimates behave? The MM estimates are minimum flows to account for change in migrant stocks. Each annual flow will be estimated to be approximately one fifth of the quinquennial flow estimate.

$$\hat{M}^A_1 \approx \hat{M}^A_2 \approx \cdots \approx \hat{M}^A_5 \approx \frac{1}{5}\hat{M}^A_{1-5}. \tag{10}$$

In contrast, the independence estimates depend on the magnitude of migrant stocks, which we assume to be approximately constant over each five-year period. Rather than each

annual estimate being roughly one fifth of the quinquennial estimate, each annual estimate under the independence structure will be nearly the same as the quinquennial estimate:

$$\hat{M}_1^I \approx \hat{M}_2^I \approx \cdots \approx \hat{M}_5^I \approx \hat{M}_{1-5}^I. \tag{11}$$

Substituting these expressions into Equation 7 allows us to rewrite our minimization in terms of annual quantities rather than quinquennial quantities. The minimization problem from Equation 7 can be expressed approximately as:

$$\text{minimize}_{w\in[0,1]} \ \|5M_1 - (5w\hat{M}_1^A + (1-w)\hat{M}_1^I)\|_F^2, \tag{12}$$

or, equivalently,

$$\text{minimize}_{w\in[0,1]} \ \left\| M_1 - \left(w\hat{M}_1^A + (1-w)\left(\frac{1}{5}\hat{M}_1^I\right)\right)\right\|_F^2. \tag{13}$$

This optimization problem now expresses a value of w which minimizes expected loss for five-year estimates in terms of only annual flows. To choose a w which works well for five-year data when we have annual data available, we substitute annual flow data along with annual MM flow estimates and independence-structured annual flow estimates into Equation 13. The minimization problem reduces to a quadratic equation in w, which is easily solved.

Altogether, we estimated the optimal value of w separately using four different combinations of ground-truth data and optimization criterion. In all cases we choose w to minimize mean squared error in migration rates, but we test two definitions of migration rates which differ only in the population in the denominator. The four combinations of data and criterion are:

1. In-flows to OECD countries. Denominator on migration "rates"[1] is the population of the destination country.

2. Out-flows from OECD countries. Denominator on migration rates is the population of the origin country.

3. IMEM flows. Denominator on migration rates is the population of the origin country.

4. IMEM flows. Denominator on migration "rates" is the population of the destination country.

Figure 6 plots optimal values of w, broken down by year as well as data and optimization criterion. Values close to 1 in all instances indicate that the number of non-movers is nearly maximized regardless of selected criterion. As such, our model still incorporates the desirable interpretation from the MM model that migrating is in some sense more costly than staying in one place, although we no longer force the number of non-movers to be as large as possible. Furthermore, optimal w values are fairly stable across both data sources and time. Slightly lower values of w for IMEM are consistent with relatively low barriers to migration within Europe. The left column of Table S1 compares average optimal annual values of w across

[1]The "rates" in definitions 1 and 4 do not meet the proper definition of a rate, as described below.

data sources. The third column extends these to five-year estimates using the method from Equation 13, while the fourth column additionally incorporates a correction for individuals who move multiple times within a five-year period, described in the section below.

Final five-year migration estimates reported in this analysis use the value of $w = 0.8704$ based on IMEM estimates of flows within Europe, and taking populations of origin countries as denominators in the definition of migration flow rates. Although any of the values in the right column of Table S1 is defensible, we chose this value based on two considerations. Firstly, we prefer to choose a value of w which optimizes estimated flow rates using a standard definition of "rate" with the population at risk of migration in the denominator. Secondly, the IMEM data are of better quality than the OECD data in at least two senses. Namely, they use a unified definition for what constitutes a migration event, and they correct for the underestimation inherent in flow estimates derived from administrative records.

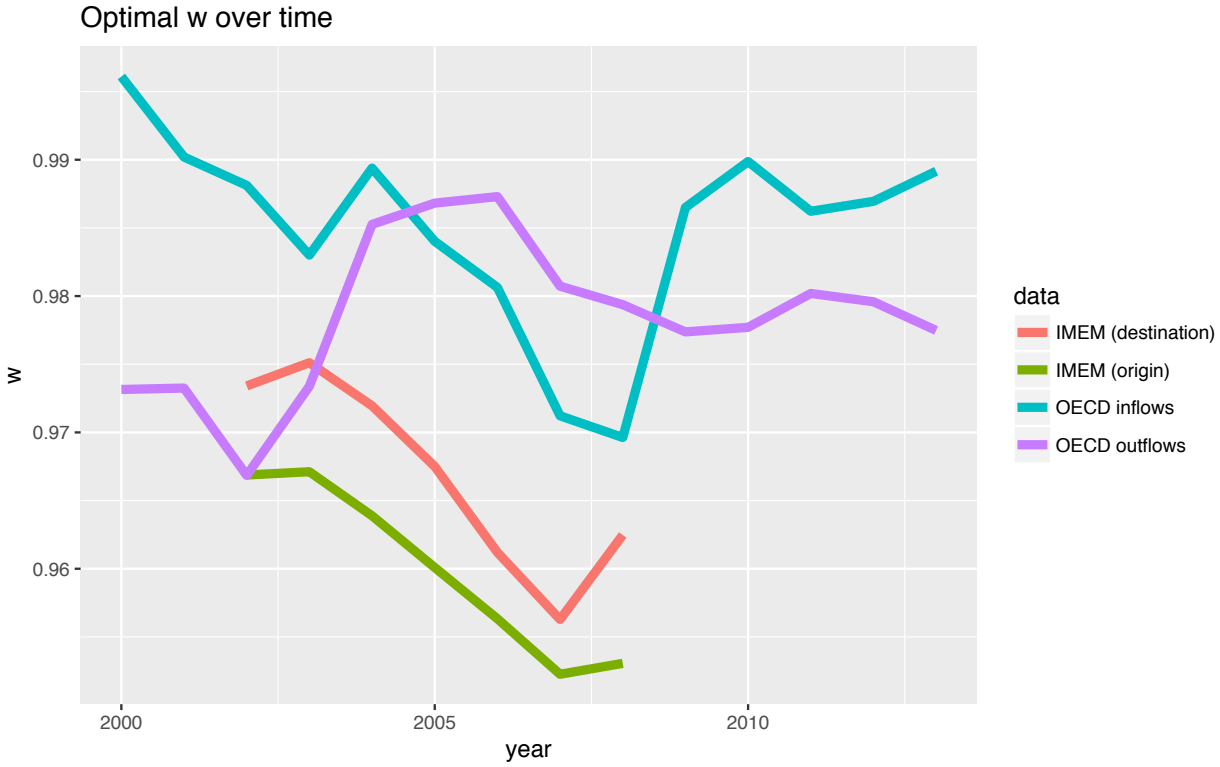

Figure 6: One-year estimates of the optimal value of w. Optimal values are fairly stable over time. Values are also similar across North-South flows (OECD out-flows), South-North flows (OECD in-flows), and North-North flows (IMEM).

2.4.1 Relaxing the one-year/five-year assumption

We now revisit the assumption made above that the number of five-year transitions is equal to the sum of the one-year transitions. In studies on internal migration in Western countries [11, 9, 12], this assumption has been found to be substantially violated, due to individuals migrating more than once within a given time period. Given an empirical estimate of the

Table S1: Estimated optimal values of w. Bolded value based on estimated optimal five-year weight for OECD outflows was the value selected to produce final reported five-year estimates. The LB-corrected value refers to the Long-Boertlein index of repeat migration.

Data used for estimation	Annual w	Five-year w	Five-year LB-corrected w
OECD Inflows only	0.9846	0.9050	0.9930
OECD Outflows only	0.9793	0.8909	0.9335
IMEM flows (denominator: origin pop.)	0.9597	0.7884	**0.8704**
IMEM flows (denominator: destination pop.)	0.9666	0.8126	0.9126

extent to which this assumption is violated, we can translate that assumption into a modified procedure for estimating a five-year optimal value of w from annual data. One means of quantifying the discrepancy between one-year and five-year transitions is the Long-Boertlein index of repeat migration [13]. This index, which we denote by ℓ, is defined as five times the one-year flow rate divided by the five-year flow rate. An index of $\ell = 1.5$ indicates that for each person who lived in a different location at the start and end of the five-year period, there were actually a total of 1.5 migration events during the period. Equivalently, an index of 1.5 indicates that the true five-year flow rate is only 3.33 times as large as the sum of the one-year transitions.

The literature provides few numerical estimates provided for typical values of ℓ. Rogers, Raymer, and Newbold [12] estimated internal values for a Long-Boertlein index of repeat migration between US regions, finding values of 1.41, 1.54, and 1.66 for different five-year time periods. Newbold [11] performed a similar study on Canadian regions, estimating values of 1.51 for a five-region breakdown of Canada, and 1.50 for a nine-region breakdown. Although these estimates are all based on internal rather than international migration, we are unaware of published international estimates. Consequently, we take an estimated value of $\ell = 1.5$ in our work.

Given a value for ℓ, we can adjust the expressions used to estimate the optimal value of w for minimizing loss on five-year estimates. Previously, we had made an assumption when estimating w that each one-year flow was approximately equal to one fifth of the five-year flow. Given a Long-Boertlein index of ℓ, we can modify equation 9 to

$$M_1 \approx M_2 \approx \cdots \approx M_5 \approx \frac{\ell}{5} M_{1-5}. \tag{14}$$

Carrying this approximation through the subsequent calculations, the procedure for estimating a five-year w now reduces to the minimization problem

$$\text{minimize}_{w \in [0,1]} \left\| \frac{1}{\ell} M_1 - \left(w \hat{M}_1^A + (1-w) \left(\frac{1}{5} \hat{M}_1^I \right) \right) \right\|_F^2, \tag{15}$$

which differs from the previous minimization problem only in the factor of $1/\ell$. This minimization problem is quadratic in w and can be solved analytically. Solving this problem with the value $\ell = 1.5$ produces the estimates for the optimal value of w in the rightmost column of Table S1.

3 Assessing uncertainty in flow estimates

Although the main product of our work is estimated flows m_{ijkt}, the degree of uncertainty in those estimates is also of interest. We have taken a simple approach to quantifying uncertainty in our estimates by examining the discrepancies between the pseudo-Bayes estimates of annual flows and the 'bronze standard' provided by the IMEM median estimates. The basic principle here is that we estimate a distribution for the observed residuals between the pseudo-Bayes and IMEM estimates, and the variance of that distribution will encapsulate the degree of uncertainty in our estimates. Figure 7 is a smoothed scatterplot comparing the IMEM flow estimates (in raw number of individuals) to the pseudo-Bayes estimates produced using the annual w of 0.9597 from Table S1. This plot is on a log-log scale with an additional constant of 1 added to each estimate to dampen the influence of estimates smaller than a single individual. In this figure, we see a strong linear correspondence between the logged IMEM and pseudo-Bayes estimates ($R^2 = 0.90$).

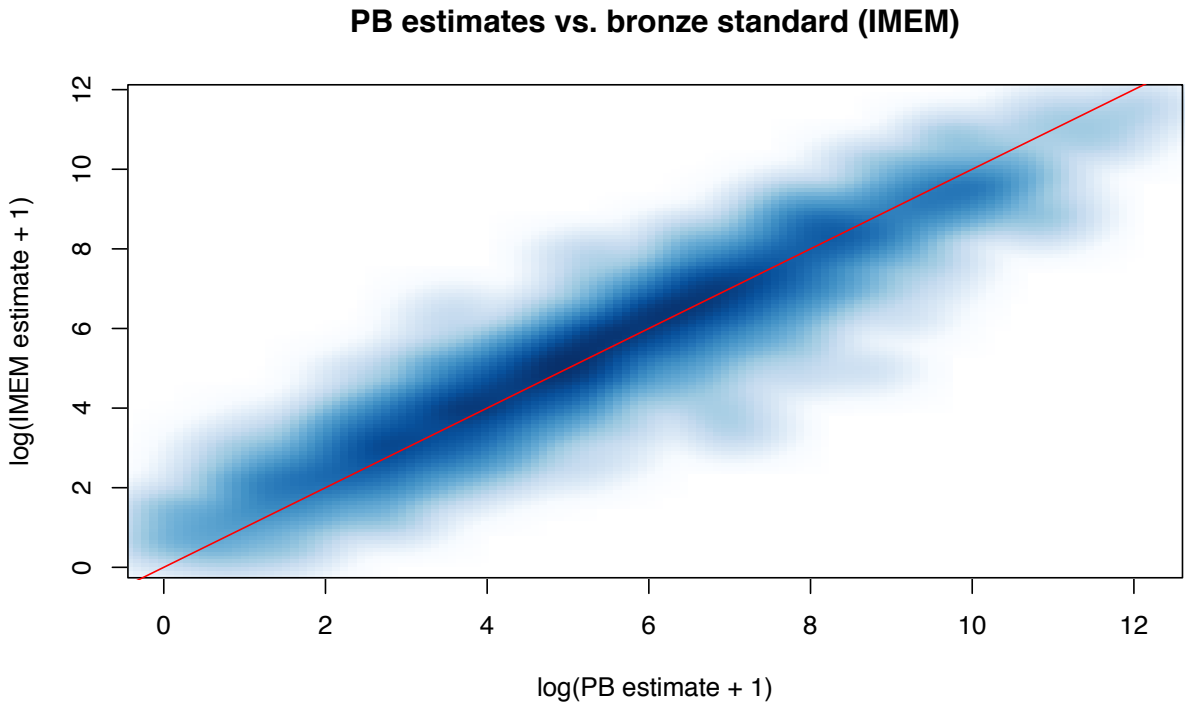

Figure 7: Smoothed scatter plot comparing log(IMEM flow estimates + 1) to log(pseudo-Bayes flow estimates + 1). Darker colors indicate a higher density of points in the scatter plot. Red line is the $y = x$ line.

To determine the degree of uncertainty in our estimates, we fit the following error model. Define y_{ijt} as $\log(\hat{m}_{ijt}^{(IMEM)} + 1)$, the logged IMEM median estimate of the total flow from country i to country j during year t (with an added constant of 1). Similarly, define x_{ijt} as

$\log(\hat{m}_{ijt}^{(PB)} + 1)$. We assume

$$y_{ijt} = \beta_0 + \beta_1 \cdot x_{ijt} + \varepsilon_{ijt} \tag{16}$$

$$\varepsilon_{ijt} \overset{ind}{\sim} t(\text{mean} = 0, \text{scale} = \sigma, \text{df} = \nu). \tag{17}$$

Under this model, the estimated values of the t distribution's scale factor and degrees of freedom control the width of a confidence interval, while the values of β_0 and β_1 control its center. From the data, we estimated a best-fit t distribution with 7 degrees of freedom and scaling factor of $\sigma = 0.59$, which we used to produce our uncertainty estimates. For annual flows of at least 1,000 individuals, this roughly corresponds to an 80% confidence interval whose lower and upper bounds differ by a factor of 5.4.

However, in applying this model to the global data, we *do not* carry over the same estimates of β_0 and β_1, which control the linear bias in our model. This bias is relatively mild in Europe, with most of the data hewing fairly closely to the $y = x$ line. One issue with carrying over the bias model from the IMEM data to the rest of the world is the problem of extrapolating to flow estimates that are larger in magnitude than those used to fit the model, which is problematic as some areas outside Europe regularly have larger flow volume than anything observed within Europe. Instead, we make the optimistic assumption that our global estimates are unbiased, but that the degree of uncertainty in Europe is representative of the degree of uncertainty elsewhere. That is, in producing confidence intervals globally, we have reduced our model to

$$y_{ijt} = x_{ijt} + \varepsilon_{ijt} \tag{18}$$

$$\varepsilon_{ijt} \overset{ind}{\sim} t(\text{mean} = 0, \text{scale} = 0.59, \text{df} = 7). \tag{19}$$

This may result in undercoverage of our confidence intervals if this assumption proves to be unsuitable.

Two further wrinkles are discussed in the following subsections. The first is the applicability of this error model to other levels of aggregation. The second is a further adjustment to account for between-year correlations in errors.

3.1 Error models at other levels of aggregation

It is worth noting that the IMEM data used as our point of comparison are at the lowest possible level of aggregation, but are summed over the place of birth index, k. Results in the main paper assume that error modeling at this level of aggregation also appropriately captures the amount of variability at the fully disaggregated level, allowing this same error model to be applied to flow estimates including place of birth, as occurs in both our table of largest flows and the USA/Mexico case study.

However, this model does not appear to be suitable for capturing uncertainty when rolled up to higher levels of aggregation (for example, uncertainty in regional aggregates or the total global migration flow.) A particular issue is that the t error model has long tails, and when summing up many t-distributed variables, the behavior of the sum is strongly influenced by the possibility that one or more observations will land in the long right tail. More work would be necessary to extend the current error model to one which produces reasonable

intervals for aggregated flow estimates, but this should not affect the suitability of the error model for individual flows considered marginally.

3.2 Extending the annual error model to a five-year error model

The raw error model which we fit to IMEM data above is suitable for producing confidence intervals in estimates of annual flows, but our main results are on five-year rather than annual flows. In analysis on the annual data, we find that the discrepancy between the PB estimates and the IMEM estimates exhibits strong temporal correlations. However, since they are not perfectly correlated, the confidence interval for the sum of five annual flows is narrower than five times the width of an annual interval. Intuitively, this narrowing occurs because while perfectly correlated errors will always reinforce one another, imperfectly correlated errors will on average contain some high values and some low values, with the sum tending to be more moderate. With that in mind, our general procedure here will be to estimate the temporal autocorrelation in error terms, and derive from that autocorrelation an estimate of how much a five-year confidence interval should be narrowed relative to five times the width of a one-year confidence interval.

Our marginal error model looked like

$$\log(\hat{m}_{ijt}^{(PB)} + 1) = \log(\hat{m}_{ijt}^{(IMEM)} + 1) + \varepsilon_{ijt}, \tag{20}$$

with $\varepsilon_{ijt} \sim t(\text{mean} = 0, \text{scale} = 0.59, \text{df} = 7)$. For the remainder of this analysis, we will make two simplifying approximations. Firstly, for computational tractability, we will replace the t error model with a normal error model, with a variance of $\sigma^2 = 0.63$ estimated by fitting a normal model to the data. Secondly, we will drop the $+1$ from within the logs, since $\log(x+1) \approx \log(x)$ so long as x is large, and the quantities of most interest to us are the largest migration flows.

We now extend this to a model with autocorrelated errors, of the form

$$\boldsymbol{\varepsilon} := \begin{pmatrix} \varepsilon_{ijt} \\ \varepsilon_{ijt+1} \\ \varepsilon_{ijt+2} \\ \varepsilon_{ijt+3} \\ \varepsilon_{ijt+4} \end{pmatrix} \sim MVN \left(\mathbf{0}, \sigma^2 \cdot \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \right). \tag{21}$$

The question of interest is really what is the distribution of the quinquennial error given by $\mathbf{1} \cdot \boldsymbol{\varepsilon}$ and how does it compare to the distribution of ε_{ijt}.

Now note that our estimated five-year flow is comprised of a sum of five correlated log-normal random variables.[2]

We now consider the distribution of this sum. A commonly used approximation is that sums of log-normal random variables are nearly log-normal, and we now carry through that approximation to determine the distribution of the sum. We carry out a method-of-moments style estimate by calculating only the first two moments of our sum of correlated log-normals,

[2]The Long-Boertlein correction discussed earlier could also be present in these expressions as a constant term. For simplicity, we ignore it here because the effects of a constant scaling factor cancel out in the final analysis.

and assuming that the sum itself approximately follows a log-normal distribution with those same first two moments.

In the general case, assume that we have a multivariate random vector of the form

$$(X_1, \ldots, X_n)' \sim MVN(\boldsymbol{\mu}, \Sigma), \tag{22}$$

and we want to know the first two moments of

$$\sum_{i=1}^{n} k_i \cdot \exp(X_i). \tag{23}$$

It's fairly straightforward to derive expressions for those first two moments. They are

$$\mathrm{E}[\sum_{i=1}^{n} k_i \exp(X_i)] = \sum_{i=1}^{n} k_i \exp(\mu_i + \sigma_i^2/2) \tag{24}$$

and

$$\mathrm{E}[(\sum_{i=1}^{n} k_i \exp(X_i))^2] = \left(\sum_{i=1}^{n} k_i^2 \exp(2\mu_i + 2\sigma_i^2)\right) + \left(\sum_{i \neq j} k_i k_j \exp(\mu_i + \mu_j + \rho_{ij}\sigma_i\sigma_j + \frac{\sigma_i^2 + \sigma_j^2}{2}).\right) \tag{25}$$

Returning to our application, many of the terms in this expression can be simplified. Our error model is constrained such that the μ_i's are all zero, the σ_i^2's are all equal, and the ρ_{ij} terms are given by $\rho^{|i-j|}$. We also need to make a further substantive assumption to get to useful expressions, about the relative values of the k_i's. In our application, the k_i's represent the true annual migration flows, m_t^a. We'll assume a regime where each annual flow is constant across a five-year period, which approximately holds for the IMEM data. Under that assumption, the k_i terms can be replaced by just k. Then the first two moments simplify greatly. The first moment simplifies to

$$nk \exp(\sigma^2/2) \tag{26}$$

and the second moment simplifies to

$$nk^2 \exp(2\sigma^2) + 2k^2 \exp(\sigma^2) \sum_{i \neq j} \exp(\rho^{|i-j|}\sigma^2). \tag{27}$$

This form is now simple enough to perform numerical analysis. For any given values of n, k, σ^2, and ρ, we can compute the width of an 80% interval for the distribution of the n-year aggregate and compared it to n times the interval width for the one-year flows. (We call this an "adjusted" ratio, because of the extra factor of n as compared to the raw ratio of interval widths.) That adjusted ratio should be 1 if the errors are perfectly correlated. In the normal case, the adjusted ratio would be $1/\sqrt{n}$ for $\rho = 0$, regardless of the values of k and σ^2. In the log-normal case, however, the adjusted ratio depends on all three of n, σ^2, and ρ.

Returning to our migration context, we estimated $\sigma^2 = 0.63$ and $\rho = 0.92$ via analysis of the annual IMEM data, leading to an adjusted ratio of 0.958. That is to say, we narrow our five-year intervals by a factor of 0.958 (relative to five times the width of a one-year interval) to account for the fact that the one-year log-normal errors are less than perfectly correlated. This narrowing is a separate adjustment from the Long-Boertlein correction for the one-year/five-year problem, which is also present.

4 Bayes and pseudo-Bayes estimators for entries in contingency tables

This section outlines the details of standard Bayes and pseudo-Bayes estimators for multinomial probabilities, as well as how our estimator of migration counts differs from the standard estimators. Mathematical identities in this section are primarily drawn from Chapter 12 of Bishop, Fienberg, and Holland's *Discrete Multivariate Analysis: Theory and Practice* [14].

4.1 Standard Bayes and pseudo-Bayes estimators for multinomial cell probabilities

A typical estimation scenario for multinomial tables is that the observations consist of a complete matrix of cell counts, X, with entries $\{x_{ij}\}$, which sum to a grand total, n. Cell counts are assumed to be drawn from a multinomial distribution. The quantity to be estimated is the true underlying cell probability matrix, P, with entries $\{p_{ij}\}$.

In the presence of an assumed prior distribution for the cell probabilities, P, we can derive a posterior distribution for P given the observed cell counts, X. A standard choice is to assume a Dirichlet distribution on P with parameter matrix B, with entries $\{b_{ij}\}$. It is also common to reparameterize the Dirichlet distribution into a prior mean matrix, A, and an overall concentration parameter, K, that controls how tight the prior distribution is about its mean. This alternative parameterization is related to the original parameters, B, by the equations

$$K := \sum_{ij} b_{ij} \tag{28}$$

and

$$A := \frac{B}{K}. \tag{29}$$

Under this decomposition, the prior mean for p_{ij} is given by a_{ij}.

Then we can express the posterior mean for P in the following form:

$$\mathrm{E}[P|B, X] = \frac{n}{n+K}(X/n) + \frac{K}{n+K}A. \tag{30}$$

That is, the Bayes estimator for P is a convex combination of the sample cell probabilities and the prior probabilities.

The *bona fide* Bayes estimator can then be extended to a pseudo-Bayes estimator in a number of ways. Note that a true Bayes estimator requires a prior belief about cell probabilities to be specified in advance, while it may be preferable in practice to allow these prior probabilities to be influenced by observed data.

One possible extension is to apply some additional structure to the prior mean matrix, A. A common empirical choice of prior recommended by [14] is to choose $A_{ij} \propto X_{i+}X_{+j}$. This is the independence-structured prior which produces the expected cell counts $\hat{M}^I$ which appear in our estimator.

Furthermore, in a genuine Bayes estimator, the concentration parameter K is a property of the prior distribution on P. However, one may prefer to specify only the prior means,

A, while choosing K in a way which minimizes the expectation of some loss function. The value of K which minimizes Bayes risk under squared error loss can be shown to be

$$K(P, A) = \frac{1 - \|P\|_F^2}{\|P - A\|_F^2}. \tag{31}$$

Of course, this optimal value for K depends on the unknown value of P. Instead, one might estimate the optimal value of K with

$$\hat{K}(\hat{P}, A) = \frac{1 - \|\hat{P}\|_F^2}{\|\hat{P} - A\|_F^2}, \tag{32}$$

where $\hat{P}$ is an initial estimate of P. This gives a pseudo-Bayes estimator which often lowers risk in practice.

4.2 Our pseudo-Bayes estimator

Again, the true Bayes estimator for cell probabilities in a contingency table with known entries X is given by

$$\frac{n}{n + K}(X/N) + \frac{K}{n + K}A, \tag{33}$$

or equivalently,

$$w(X/n) + (1 - w)A. \tag{34}$$

The typical extension to a pseudo-Bayes estimator is to allow K (or equivalently w) and A to depend on the data.

However, in our migration application we introduce two further modifications. Firstly, we have an alternate route to estimating w. Instead of leveraging an initial set of estimates into a guess at a better value for w, we choose a value for w by validating against external data sources (i.e. OECD and IMEM flow estimates). The motivation is still minimization of expected loss, but we accomplish this via validation against 'bronze standard' data rather than analytical results. Secondly, we don't have a set of observed cell counts X. What we do have are estimated cell counts $\hat{M}^A$ produced using Abel's minimum migration method [1].

Our pseudo-Bayes estimator for cell probabilities takes the form

$$w(\hat{M}^A/n) + (1 - w)(\hat{M}^I/n). \tag{35}$$

Taking the place of observed cell counts, X, are the MM estimates of cell counts, $\hat{M}^A$. In place of prior cell probabilities, A, we use an empirical prior, $\hat{M}^I/n$, in which the matrix of prior probabilities has rank one, which is equivalent to an assumption of independence between rows and columns. Finally, w is a weight estimated by minimizing some loss function on OECD and IMEM data.

There is some cost associated to introducing these modifications to the Bayes estimator for cell probabities. Most Bayes estimators have the property of converging to some sample statistic asymptotically, which is a desirable feature because the sample statistic is usually a

consistent estimator of the quantity of interest. That is not the case here. As n increases, our weight w doesn't change because we estimated w from an external data source. However, the MM estimator $\hat{M}^A$ is itself asymptotically biased, so asymptotic convergence to $\hat{M}^A$ would not guarantee consistency anyhow.

While our estimator fits into the morphological family of pseudo-Bayes estimators, it's worth keeping in mind that the motivation for using it is somewhat different from the typical Bayes estimator. The scenario is not one of trying to adjust a consistent estimator in a way that lowers Bayes risk. Nor are we improving a low-sample-size estimate by incorporating prior information. Instead, we're making up for a structural weakness in the MM estimator (namely, that no cross-flows exist) by shrinking towards a prior mean which lacks that weakness.

References

[1] Abel GJ (2013) Estimating global migration flow tables using place of birth data. *Demographic Research* 28:505–546.

[2] Zipf GK (1946) The p1 p2/d hypothesis: on the intercity movement of persons. *American Sociological Review* 11(6):677–686.

[3] Lee ES (1966) A theory of migration. *Demography* 3:47–57.

[4] Sjaastad LA (1970) The costs and returns of human migration in *Regional Economics*. (Springer), pp. 115–133.

[5] Kim K, Cohen JE (2010) Determinants of international migration flows to and from industrialized countries: A panel data approach beyond gravity. *International Migration Review* 44(4):899–932.

[6] Abel GJ (2017) Estimates of global bilateral migration flows by gender between 1960 and 2015. *International Migration Review*.

[7] Organization for Economic Co-Operation and Development (2015) *OECD International Migration Database*. (Organization for Economic Co-operation and Development, Paris).

[8] Raymer J, Wiśniowski A, Forster JJ, Smith PW, Bijak J (2013) Integrated modeling of european migration. *Journal of the American Statistical Association* 108(503):801–819.

[9] Rees PH (1977) The measurement of migration, from census data and other sources. *Environment and Planning A* 9(3):247–272.

[10] Kitsul P, Philipov D (1981) The one year/five year migration problem. *Advances in Multiregional Demography* pp. 1–34.

[11] Newbold KB (2005) Spatial scale, return and onward migration, and the Long-Boertlein index of repeat migration. *Papers in Regional Science* 84(2):281–290.

[12] Rogers A, Raymer J, Newbold KB (2003) Reconciling and translating migration data collected over time intervals of differing widths. *The Annals of Regional Science* 37(4):581–601.

[13] Long JF, Boertlein CG (1990) Comparing migration measures having different intervals., (U.S. Department of Commerce, Bureau of the Census, Washington, DC), Technical report.

[14] Bishop YM, Fienberg SE, Holland PW (1975) *Discrete Multivariate Analysis: Theory and Practice*. (The MIT Press).