Estimation of flows within the intra-EU migration matrix

Michel POULAIN and Luc DAL
GéDAP-UCL,
Revised version
30 June 2008

Objective of the estimation procedure

On the basis of data provided by countries and considering reasonable and simple assumptions we will propose estimated value for each migration flow occurring between pairs of EU member States.

Year of observation: 2004 (the same methodology is applicable using 2002, 2003, 2005 and 2006 data)

Countries involved: All 27 EU MS except Malta, where no data at all are available, plus Norway and Switzerland. Accordingly the size of the migration matrix will be 28 X 28. Considering MIMOSA countries, only Iceland, Liechtenstein and Malta are not included.

Data used

The requested migration data is immigration by country of previous residence and emigration by country of next residence. The 28 countries are classified in two groups as follows:

- A first group of countries (13) providing consistent migration data for 2004, these countries will be called **referee countries**.
- Among the remaining 15 **non-referee countries** where we consider first countries **with incomplete or not consistent data** (7 countries). Concretely for Italy, no data was available for 2004 and data from 2003 have been used. For Belgium, no data is available and GéDAP UCL proposes estimations based on official migration records. For UK, the International Passenger Survey is based on a sample so that we have no migration originated from some small countries as Luxembourg, Latvia… Data from previous years will be used for these countries. For migrations between UK and Ireland, some past estimation is proposed. For Switzerland the data on migration flows by country of citizenship is considered as proxy for the country of next or previous residence. And finally we consider the **non-referee countries where no data exists** or is considered as unreliable and where the population stock by country of citizenship will be used as proxy (8 countries). The most recent distribution of the population stock by country of citizenship will be used mostly based on the last census that was part of the 2000 round censuses

Assumptions

Compared to the UN definition that considers an intended duration of stay abroad or in the country of one year or more, most countries record migration events with a shorter time criterion (three or six months) or no time criterion at all. Doing so the countries concerned would register more migrations than what is expected following the UN recommendations.

However the main problem in terms of reliability is linked to the fact that international migrations have generally to be self-reported by the persons concerned. Accordingly a relatively large under coverage is observed and the level of that varies between countries depending if the administrative rules for registration are strictly followed or not.

In the proposed methodology, we will assume that the **level of under coverage or the relative gap is constant for a given country whatever the EU country of destination for emigration and the EU country of origin for immigration.** Under this assumption specific **correction multiplicative factors** may be calculated for each country and applied to all immigrations or all emigrations registered by that given country in order to estimate the real migration flows.

For countries not having exactly the requested data we will use the data from previous years or the data of immigrations and emigrations by country of citizenship. Finally when no data is available we will use the stock data by country of citizenship at the last available census. In order to support these choices we may prove that there is a large correlation between figures from two successive years, between migration figures by country of origin and destination and migration by country of citizenship and also between stock data by country of citizenship and migration flows by country of origin or destination. According to this methodology the correction factors cannot be compared between countries where different types of data are used. Such comparison is only possible between the 13 referee countries and it includes the impact of both a different time criterion and a different level of coverage. Only for countries using the same definition for the same observation year the correction factors may indicate the relative level of under coverage between these two countries.

Finally in order to fix concretely the level of correction it is assumed that **Sweden is correctly recording all international immigrations following the UN recommendations**. This is plausible as it is generally admitted that Nordic countries are the one with the highest coverage and between these Nordic countries, Sweden is the only one using the 12 months criterion for defining international immigrant.

Methodology

We consider two matrices with migration data as follows:

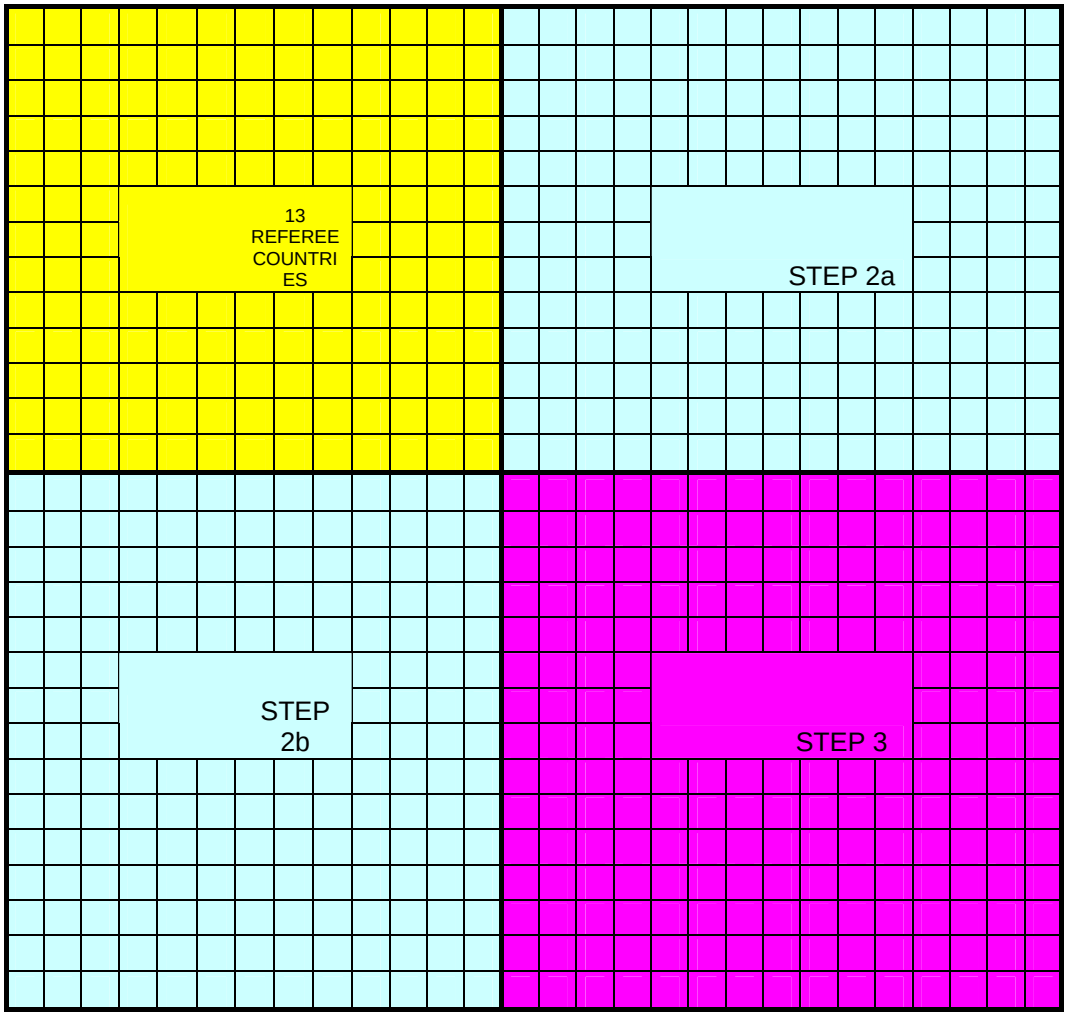
- I is the immigration matrix of dimension n where countries of immigration (j) are proposed as head of columns and country of origin (i) as heads of rows.
- E is the emigration matrix of dimension n where countries of emigration (i) are proposed as head of rows and countries of destination as heads of column (j).

The two matrices are directly comparable as in both the origin country is always head of row and the destination country head of column. When a migration flow is zero, we will put one unique migration in order to make the proposed method operational.

We will develop the methodology in three successive steps based on the identification of two groups of countries (figure 1). The first step is based on data produced by the 13 so-called **referee countries** and 13 correction factors will be estimated for immigrations recorded by these countries and 13 others for emigrations recorded by these countries The second step will consider the flows between the 13 referee countries and the 15 non

referee countries and two specific correction factors will be estimated by considering the corrected migration flows proposed by the referee countries and the proposed figures by the second group of countries. The correction factors estimated in the second step for the 15 remaining countries will allow estimating the migration flows occurring between these 15 remaining countries.

Figure 1. The estimation procedure is divided in three steps by considering the two groups of countries, the 13 referee countries and the 15 non-referee countries.



We will estimate two series of correction factors α_j and β_i so that $\alpha_j * I_{i,j} = \beta_i * E_{i,j}, \forall i = 1..n, \forall j = 1..n$.

The factors α_j and β_i are respectively the correction factors for immigrations in country j and emigrations from country i. For each migration flow between countries i and j, the final estimation would be $S_{i,j} = \frac{\alpha_j * I_{i,j} + \beta_i * E_{i,j}}{2}$.

The estimation will result from the minimization of the weighted sum of the square of differences between the corrected immigration figure $\alpha_j * I_{i,j}$ and the corrected emigration figure $\beta_i * E_{i,j}$.

The weight will be similar to the one appearing in the χ^2 distance: $\forall i = 1..n, \forall j = 1..n : \rho_{i,j} = I_{i,j} + E_{i,j}$

In order to minimize the following sum $\sum_{i,j} \frac{\left(\alpha_j * I_{i,j} - \beta_i * E_{i,j}\right)^2}{\rho_{i,j}}$ we shall derive it by each of the 2n variables and doing so we will obtain a system of 2n equations with 2n unknowns.

However this system is undetermined (it accepts a infinity of solution defined with a variable constant multiplicative value).

In order to fix the final value for each variable, we will impose an additional constraint to the system. **Therefore we will assume that the correction factor for immigrations registered in Sweden is equal to 1**. Doing so we consider that like other Nordic countries, Sweden presents a high level of coverage and moreover the time criterion used is 12 months what is requested by the UN recommendations. As the system of equations to be solved with that additional constraint has been found very unstable we will first introduce the constraint that the sum of estimated flows is equal to the sum of the maximum between immigration and emigration figures for each flow. Thereafter we will transform proportionally every correction factors so that α_{SE} equal 1. Please note that under this additional constraint, the total number of migrations between the concerned countries (immigration or emigration as both total figures will be the same) will be higher than the sum calculated in the two original matrices.

The problem is therefore a constrained optimization of 2n unknown.

The *Lagrangian Multiplicator* method allows solving this problem, by introducing the constraint in the expression to minimize:

$$\sum_{i,j} \frac{\left(\alpha_j * I_{i,j} - \beta_i * E_{i,j}\right)^2}{\rho_{i,j}} - \lambda\left(\sum_{i,j}\left(\frac{\alpha_j * I_{i,j} + \beta_i * E_{i,j}}{2} - Max\left(I_{i,j}, E_{i,j}\right)\right)\right)$$ where λ is unknown.

Taking partial derivates with respect of all parameters we obtain:

$$\begin{cases} \forall j = 1..n : \frac{\partial}{\partial \alpha_j} = \sum_i \left(\frac{2I_{i,j}\left(\alpha_j I_{i,j} - \beta_i E_{i,j}\right)}{\rho_{i,j}} - \frac{\lambda I_{i,j}}{2} \right) \\ \forall i = 1..n : \frac{\partial}{\partial \beta_i} = \sum_j \left(\frac{-2E_{i,j}\left(\alpha_j I_{i,j} - \beta_i E_{i,j}\right)}{\rho_{i,j}} - \frac{\lambda E_{i,j}}{2} \right) \\ \frac{\partial}{\partial \lambda} = -\sum_{i,j} \left(\frac{\alpha_j I_{i,j} + \beta_i E_{i,j}}{2} - Max\left(I_{i,j}, E_{i,j}\right) \right) \end{cases}$$

Equating each member to 0, the optimization problem becomes now a simple (2n+1)*(2n+1) system.

$$\begin{cases} \forall j = 1..n : \dfrac{\partial}{\partial \alpha_j} = 0 \\ \forall i = 1..n : \dfrac{\partial}{\partial \beta_i} = 0 \\ \dfrac{\partial}{\partial \lambda} = 0 \end{cases}$$

It's linear without any constraint, and it is easy to solve it.

$$\begin{cases} \forall j = 1..n : \sum_i \left(I_{i,j} \left(\alpha_j I_{i,j} - \beta_i E_{i,j} \right) - \dfrac{\lambda I_{i,j} \rho_{i,j}}{4} \right) = 0 \\ \forall i = 1..n : \sum_j \left(E_{i,j} \left(\alpha_j I_{i,j} - \beta_i E_{i,j} \right) + \dfrac{\lambda E_{i,j} \rho_{i,j}}{4} \right) = 0 \\ \sum_{i,j} \left(\alpha_j I_{i,j} + \beta_i E_{i,j} \right) = 2 \sum_{i,j} \left(Max \left(I_{i,j}, E_{i,j} \right) \right) \end{cases}$$

First step: Application of this methodology to the 13 referee countries

The results for the so-called referee countries are presented in table 1 where the correction factor have been divided by an appropriate constant so that α_{SE} equal to 1.

Table 1. Correction factors estimated for the 13 referee countries during the first step of the procedure.

Country	Correction factor for immigration	Correction factor for emigration
DK	0,865	0,904
FI	1,132	1,055
SE	1,000	1,079
NO	1,000	0,999
BE	1,026	1,211
CZ	4,120	6,803
DE	0,705	0,823
ES	0,979	5,216
LV	3,628	5,752
LT	2,959	2,300
LU	0,991	1,194
NL	0,988	1,043
AT	1,039	1,694

A quick look on these data shows that immigration correction factors are generally lower than the corresponding emigration correction factors. The lowest value concerns immigrations recorded in Germany; thereafter emigrations recorded by Germany and immigrations and emigrations recorded by Denmark. The highest values are found for the Czech Republic, Latvia and Lithuania for both immigrations and emigrations and also for emigration in Spain. All these results were expected following the numerous investigations done in the THESIM project.

Why do we propose these 13 countries as referee countries excluding other ones like Italy, UK that also provide migration flow data? And why countries like Spain, Latvia, Lithuania and the Czech Republic are not excluded? We did a lot of tests showing that any additional country will introduce a significant disturbance in the values of the parameters while if we suppress one of the 13 countries the other parameters stay relatively stable. Accordingly these 13 countries may be chosen as **referee** for the estimation procedure. In addition the variation of these parameters from year to year for 2002 and 2003 is very limited. Based on the estimated correction factors, corrected figures will be calculated for each immigration and emigration flow between the 13 referee countries. The final estimation after step one for every migration between the 13 referee countries will be the average between the two corrected figures.

The factors α_j and β_i are respectively the correction factors for immigrations into country j and emigrations from country i. For each migration flow between countries i and j, the final estimation would be $S_{i,j} = \frac{\alpha_j * I_{i,j} + \beta_i * E_{i,j}}{2}$.

The next step will estimate the migration flows between these 13 referee countries and the 15 non-referee countries.

Second step: estimation of the migration flows between the two groups of countries

For each country of the non-referee countries we will use as proxy when nothing else is available, any available migration data, reliable or less reliable, and if needed figures based on previous years and even flow or stock data by country of citizenship.

For the estimation of the correction factor for immigration we will use the estimated values based on emigrations from each country composing the 13 referee countries. $\alpha_j = \frac{\sum_i \beta_i E_{i,j}}{\sum_i I_{i,j}}$.

On a similar base the estimation of the correction factor for emigration we will use the estimated values based on immigrations in each of the 13 referee countries. $\beta_j = \frac{\sum_i \alpha_i I_{i,j}}{\sum_i E_{i,j}}$.

Table 2. Correction factors estimated for the second group of countries[1]

Country	Correction factor for immigration	Correction factor for emigration
IT	1,667	2,289
CY	0,394	2,780
PT	6,269	0,633
SI	5,053	1,850
SK	22,331	43,690
UK	0,772	1,222
CH	0,935	0,916
BG	11,176	14,912
EE	0,610	0,749
IE	0,282	0,195
GR	0,446	0,358
FR	0,168	0,163
HU	1,251	1,388
PL	4,376	7,778
RO	10,620	22,666

As explained above, these correction factors cannot be directly compared neither with the previous set of correction factors nor between themselves as each of them is related to the available specific data used for that given country. In this second step, the estimated flows will be based only on the figures provided by the referee countries. That means that in the step 2a the estimated immigration flows will be the corrected emigration flows provided by the referee countries while in the step 2b the estimated emigration flows will be corrected immigration flows provided by the referee countries. Accordingly the figures provided by the 15 non-referee countries will not be considered at all for estimation. However these figures were helpful for calculating the correction factors for the non-referee countries that will be used in the third step for estimating the migration flows between the 15 non-referee countries.

Third step: estimation of the migration flows between non-referee countries

In the last step we will use all values of the correction factors α_j and β_i as estimated in the second step and the final estimation for each migration flow will be also given by

$$S_{i,j} = \frac{\alpha_j * I_{i,j} + \beta_i * E_{i,j}}{2}$$

[1] The grey cells refer to available migration flow data while the white cells use distribution of the population by country of citizenship at the last census as proxy.