The IMEM model for estimating international migration flows between countries in Europe

James Raymer*, Jonathan J. Forster, Peter W.F. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy J. Abel

Southampton Statistical Sciences Research Institute
University of Southampton

March 1, 2011

Draft version prepared for
Population Association of America Annual Meeting 2011,
Washington, DC, March 31 - April 2

Abstract

In order to fully understand the causes and consequences of international population movements in Europe, researchers and policy makers need to overcome the limitations of the various data sources, including inconsistencies in availability, definitions and quality. In this paper, we propose a Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows. The focus is on estimating recent international migration flows between countries in the European Union (EU) and European Free Trade Association (EFTA), using data primarily collected by Eurostat and other national and international institutions. The methodology is integrated and capable of providing a synthetic data base with measures of uncertainty for international migration flows and other model parameters.

1 Introduction

IMEM (Integrated Modelling of European Migration) is a two-year project funded by NORFACE (New Opportunities for Research Funding Agency Co-operation in Europe) to develop an integrated model for estimating migration flows between countries in Europe.

In order to fully understand the causes and consequences of international population movements in Europe, researchers and policy makers need to overcome the limitations of the various data sources, including inconsistencies in availability, definitions and quality. In this paper, we propose a Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows. The focus is on estimating recent international migration flows amongst countries in the European Union (EU) and European Free Trade Association (EFTA), using data primarily collected by Eurostat and other national and international institutions. The methodology is integrated and capable of providing a synthetic data base with measures of uncertainty for international migration flows and other model parameters.

The advantages in having a consistent and reliable set of migration flows are numerous. Estimates of migration flows are needed so that governments have the means to improve their planning policies directed at supplying particular social services or at influencing

*Contact email at raymer@soton.ac.uk

levels of migration. They can also be used to inform economic models for labour market change. This is important because migration is currently (and increasingly) the major factor contributing to population change. Furthermore, our understanding of how or why populations change requires reliable information about migrants. Without this, the ability to predict, control or understand that change is limited. Finally, countries are now required to provide harmonised migration flow statistics to Eurostat as part of a new regulation passed by the European Parliament in 2007. Recognising the many obstacles with existing data, Article 9 of the Regulation states that 'As part of the statistics process, scientifically based and well documented statistical estimation methods may be used.'[1] Our proposed framework helps countries achieve this aim and provides measures of accuracy required for understanding the estimated parameters and flows.

2 Background

The reasons for international migration are many. People move for employment, family reunion or amenity reasons. Reported statistics on these flows, on the other hand, are relatively confusing or nonexistent. There are two main reasons. First, no consensus exists on what exactly is a "migration". Therefore, comparative analyses suffer from differing national views concerning who is a migrant. Second, the event of migration is rarely measured directly. Often it is inferred by a comparison of places of residence at two points in time or as a change in residence recorded by a population registration system. The challenge is compounded because countries use different methods of data collection. Migration statistics may come from administrative data, decennial population censuses or surveys. The timing criterion used to identify international migrants varies considerably between countries. For population register data, international migration may refer to persons who plan to live or have lived in a different country for a minimum period of three months, six months, one year, or even more. Research is needed to reconcile the different timings used to collect or model migration data, as well as between different collection systems.

International migration statistics also suffer from unreliability, mainly due to under-registration of migrants and data coverage (Nowok et al. 2006). This is often caused by the collection method or by non-participation of the migrants themselves. In general, migration data may be unreliable because they are often based on intentions. Emigration data are particularly problematic because migrants may not notify the population register of their movement because it is not in their interest to do so. Surveys, such as the United Kingdom's International Passenger Survey, often do not have large enough sample sizes to adequately capture the needed details for analysing migration. Without a relatively large sample size, irregularities in the data are likely to appear, such as in the country-to-country-specific flows. Furthermore, flows for certain countries may be missing for particular years or entirely. Finally, migration data may be available only for the total population, not for more detailed demographic, socioeconomic or spatial characteristics required for a particular study.

Because of all the problems associated with inconsistency and missing data, there has been a very limited amount of work carried out in the area of estimating international migration matrices. Most of the estimation work has been focused on indirect methods for particular countries, independent of others (e.g., Hill 1985; Jasso and Rosenzweig 1982; Schmertmann 1992; Van der Gaag and Van Wissen 2002; Warren and Peck 1980; Zaba 1987). There are, however, three exceptions that focus on European migration from which we can draw experiences: Poulain's (1993, 1999) 'correction factor' approach, Raymer's (2007, 2008) 'multiplicative component' approach and Brierley et al.'s (2008) Bayesian approach. The correction factor approach demonstrated the weaknesses of reported migration data and provided a simple mathematical method for adjusting the flows and making them more consistent across countries. The multiplicative component approach

[1] http://www.europarl.europa.eu/sides/getDoc.do?objRefId=140109&language=EN.

showed how standard spatial interaction models for internal migration could be applied to model international migration flows in a hierarchical manner. Finally, the Bayesian approach demonstrated the usefulness and flexibility of incorporating various forms of prior information and the importance of distributions quantifying uncertainty in the predicted values.

Recently, Raymer with colleagues at NIDI have collaborated on a Eurostat-funded project to estimate international migration stocks and flows in Europe. The work on estimating flows is described in Raymer et al. (forthcoming). The methodology adopted by the MIMOSA (MIgration MOdelling for Statistical Analyses) team represents a two-stage hierarchical procedure. The first stage harmonises the available data by using a simple optimisation procedure (Poulain 1999) benchmarked to Sweden's migration flow data, which are assumed to be measured more or less without error (see also de Beer et al. 2010). The second stage estimates the missing marginal data and associations between countries by using the available flows and covariate information. Both stages are set within a multiplicative framework for analysing migration flows. No measures of uncertainty are provided and the approach is sensitive to the model assumptions and estimation procedure.

The above works have led us to the conclusion that a Bayesian approach is the only one capable of integrating all the different types of data and expert judgements. There are two important advantages of adopting a Bayesian approach in the context of the proposed research. First, the methodology offers a coherent and probabilistic mechanism for describing various sources of uncertainty contained in the various levels of modelling. These include the migration processes, models, model parameters and expert judgements. Second, the methodology provides a formal mechanism for the inclusion of expert judgement to supplement the deficient migration data. As noted by Willekens (1994), a Bayesian approach for modelling international migration is particularly well-suited for incorporating expert judgement to substitute for data shortages. Applications of this approach in migration and population analyses include, for example, predictions of international migration from time series models (Gorbey et al. 1999; Bijak and Wiśniowski 2010), non-migratory spatial movements (Congdon 2001), forecasts of fertility (Tuljapurkar and Boe 1999) and mortality (Czado et al. 2005; Girosi and King 2008), and the estimation of population size under situations of very limited information (Daponte et al. 1999, in the study of the Kurdish population of Iraq). A thorough overview of applications of Bayesian methods in social sciences, including demographic modelling in the multistate framework, is offered by Lynch (2007).

3 Methodology

There are two key design aspects of our methodology: (1) the development of the underlying statistical framework and (2) the specification of prior information. We address each of these in turn below.

3.1 The Statistical Modelling Framework

The data of interest can be conveniently expressed in a two-way contingency table or matrix showing the origin-to-destination flows with the cell counts corresponding to the number of migrants in a specified period. Consider a matrix Z of reported migration flows (without age or sex) and a corresponding matrix Y of true migration flows with unknown entries:

$$Z = \begin{pmatrix} 0 & z_{12} & z_{13} & \dots & z_{1n} \\ z_{21} & 0 & z_{23} & \dots & z_{2n} \\ z_{31} & z_{32} & 0 & \dots & z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \dots & 0 \end{pmatrix} \quad Y = \begin{pmatrix} 0 & y_{12} & y_{13} & \dots & y_{1n} \\ y_{21} & 0 & y_{23} & \dots & y_{2n} \\ y_{31} & y_{32} & 0 & \dots & y_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & \dots & 0 \end{pmatrix}.$$

We observe counts (flows) z_{ijt}^k from country i to country j during year t reported by either the sending S or receiving R country, where $k = \{S, R\}$. We assume that z_{ijt}^k follows a Poisson distribution

$$z_{ijt}^S \sim \text{Po}(\mu_{ijt}^S), \quad \text{for all } i, j \text{ and } t \tag{1}$$
$$z_{ijt}^R \sim \text{Po}(\mu_{ijt}^R), \quad \text{for all } i, j \text{ and } t. \tag{2}$$

3.2 Measurement error model

In our model, y_{ijt} is a true flow of migration from country i to country j in year t. It includes migration flows to and from rest of world (category $i = 0$). In terms of measurement, true flows are consistent with the United Nations (UN, 1998) recommendation for long-term international migration, that is a **long-term migrant** is *a person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes his or her new country of usual residence. From the perspective of the country of departure, the person will be a long-term emigrant and from that of the country of arrival, the person will be a long-term immigrant.*

The two measurement error equations are

$$\log \mu_{ijt}^S = \log y_{ijt} + \beta_i - \log\left(1 + e^{-\kappa_i}\right) + \varepsilon_{ijt}^S, \quad \text{for all } i, j \text{ and } t \tag{3}$$
$$\log \mu_{ijt}^R = \log y_{ijt} + \gamma_j - \log\left(1 + e^{-\kappa_j}\right) + \varepsilon_{ijt}^R, \quad \text{for all } i, j \text{ and } t, \tag{4}$$

where we assume $\varepsilon_{ijt}^S \sim \mathcal{N}(0, \tau^S)$ and $\varepsilon_{ijt}^R \sim \mathcal{N}(0, \tau^R)$. The precisions of the error term do not depend on the country. Instead, they depend on whether the data are captured by sending or receiving countries as a whole. Thus we take

$$\tau^S = t_1(c), \tag{5}$$
$$\tau^R = t_2(c), \tag{6}$$

where c denotes the type of collection system (e.g., population register or survey). The number of parameters required to capture differences in accuracy, ultimately depends on our typology of collection systems, and their relative ability to capture migration flows, regardless of definition and coverage. For the moment, a model has been tested with two types of accuracy: poor and good. Good accuracy was assumed for five Scandinavian countries and the Netherlands. The accuracy is distinct for emigration and immigration.

The differences in duration of stay criterion are captured by the parameters β_i and γ_j by means of functions b and g, which are specified as

$$\beta_i = b(def(i)) = \begin{cases} \delta_1 + \log(\lambda_1) & if \quad def(i) \text{ is 0 months} \\ \delta_2 + \log(\lambda_1) & if \quad def(i) \text{ is 3 months} \\ \delta_3 + \log(\lambda_1) & if \quad def(i) \text{ is 6 months} \\ +\log(\lambda_1) & if \quad def(i) \text{ is 12 months} \\ \delta_4 + \log(\lambda_1) & if \quad def(i) \text{ is permanent} \end{cases}, \tag{7}$$

$$\gamma_j = g(def(j)) = \begin{cases} \delta_1 + \log(\lambda_2) & if \quad def(j) \text{ is 0 months} \\ \delta_2 + \log(\lambda_2) & if \quad def(j) \text{ is 3 months} \\ \delta_3 + \log(\lambda_2) & if \quad def(j) \text{ is 6 months} \\ +\log(\lambda_2) & if \quad def(j) \text{ is 12 months} \\ \delta_4 + \log(\lambda_2) & if \quad def(j) \text{ is permanent} \end{cases}. \tag{8}$$

Parameter δ_m measures the effect of a particular duration of stay definition used by country i, which is denoted by $def(i)$. The parameters were constrained so that $\delta_1 > \delta_2 > \delta_3 > 0$ and $\delta_4 < 0$ in the following way.

$$\delta_1 = d_1 + d_2 + d_3,$$
$$\delta_2 = d_2 + d_3,$$
$$\delta_3 = d_3,$$
$$\delta_4 = -d_4,$$

where $d_k > 0$ are auxiliary parameters. Parameters λ_r measure the effect of the undercount and it is assumed that $\lambda_r \in (0, 1)$.

Parameter κ_i is a country-specific random effect normally distributed

$$\kappa_i \sim \mathcal{N}(m.k_m, t.k_m),$$

where $m.k_m$ is a group-specific mean and $t.k_m$ is a group-specific precision. The logistic transformation of κ ensures that the function is bounded within a range $(0, 1)$ on the linear scale. It can be interpreted in terms of the differences in coverage with respect to the UN definition of migration. For the time being, there are three groups of coverage, that is $m = \{\text{poor}, \text{good}, \text{excellent}\}$. We further assume that $m.k_m$ is normally distributed with mean and precision hyperparameters and $t.k_m$ is gamma distributed with shape and scale hyperparameters.

For the migration to and from the rest of world there is only one equation per outflow and inflow, respectively, i.e.

$$\log \mu^S_{i0t} = \log y_{i0t} + \beta_i - \log\left(1 + e^{-\kappa_i}\right) + \varepsilon^S_{i0t}, \quad \text{for all } i \text{ and } t \tag{9}$$

$$\log \mu^R_{0jt} = \log y_{0jt} + \gamma_j - \log\left(1 + e^{-\kappa_j}\right) + \varepsilon^R_{0jt}, \quad \text{for all } j \text{ and } t, \tag{10}$$

All other parameters remain same as described above, except for β_i and γ_j, which are defined here as $\beta_i = \delta_{def(i)} + \log(\lambda_3)$ and $\gamma_j = \delta_{def(j)} + \log(\lambda_4)$.

3.3 Migration model

The true flows of migration may be modelled according to a set of covariate information. Here, we rely on migration theory and empirical evidence to drive the development of the model, see Jennissen (2004), Abel (2010) and Raymer et al. (forthcoming). The explanatory variables can be grouped into economic, demographic and geographic ones. Consider the following model of migration:

$$\begin{aligned}\log y_{ijt} = \alpha_1 + \alpha_2 \log(P_{it}) + \alpha_3 \log(P_{jt}) + \alpha_4 C_{ij} + \alpha_5 \log(T_{ijt}) + \\ \alpha_6 \log(GNI_{it}/GNI_{jt}) + \alpha_7 A_{ijt} + \alpha_8 (MS_{ij}/P_i) + \alpha_9 (MS_{ij}/P_j) + \xi_{ijt},\end{aligned} \tag{11}$$

where $\alpha = (\alpha_1, \ldots, \alpha_9)'$ is a vector of parameters. The random term ξ is assumed to be normally distributed with 0 mean and constant precision τ_y, following Brierley et al. (2008).

The following set of covariates is used:

1. The mid-year populations (averages of 1 January populations of subsequent years) in sending and receiving country, denoted as P_{it} and P_{jt}; source: NewCronos database of Eurostat.

2. Dummy variable indicating contiguity (or neighbouring countries) with 1 if countries i and j have a common border and 0 otherwise, C_{ij}; source: Mayer and Zignago (2006). Contiguity between all Scandinavian countries was assumed.

3. The ratio of the Gross National Income per capita in sending and receiving countries, GNI_{it}/GNI_{jt}. Source: World Development Indicators (2010).

4. International trade between origin and destination countries expressed as import in current USD, T_{ijt}. Source: UN Commodity Statistics Database [2].

5. Dummy variable for accession dummy A_{ijt}. It is equal to 1 for all flows between ten countries which accessed the EU in 2004 and Ireland, the United Kingdom and Sweden – countries which open their labour markets on the day of accession, for years 2004-2008.

[2] http://comtrade.un.org, accessed July 2010

6. Origin-destination migrant stocks ratios to the sending and receiving populations based on the 2000 population censuses round, MS_{ij}/P_i and MS_{ij}/P_j; source: Parsons et al. (2005).

All variables, apart from contiguity, accession and migrant stocks, were divided by their means and then logged. For Liechtenstein the following imputations were carried out:

- GNIs were assumed the same as in Switzerland (CH).
- Trade flows were taken from the statistical office's website[3], converted from CHF to USD. Import from Liechtenstein to other countries was approximated by export. Trade between LI and CH was calculated using ratios as presented in 'Liechtenstein – Industrial location' presentation[4], that is export from LI to CH to be 11% of the total and import from CH to LI to be 33% of the total.

For modelling flows to the rest of world, we use a model with additional covariates based on Raymer et al. (forthcoming).

$$\begin{aligned}\log y_{i0t} = \beta_1 + \beta_2 \log(P_{it}) + \beta_3 \log(GNI_{it}) + \beta_4 NSn_i + \beta_5 \log(MS_{0i}/P_i) + \\ \beta_6 \log(P65_{it}) + \beta_7 \log(LEW_{it}) + \xi_{i0t},\end{aligned} \tag{12}$$

and for flows from the rest of world

$$\begin{aligned}\log y_{0jt} = \beta_8 + \beta_9 \log(P_{jt}) + \beta_{10} \log(GNI_{jt}) + \beta_{11} NSn_j + \beta_{12} \log(MS_{0j}/P_j) + \\ \beta_{13} \log(P65_{jt}) + \beta_{14} \log(LEW_{jt}) + \xi_{0jt}.\end{aligned} \tag{13}$$

Errors ξ_{i0t} and ξ_{0jt} are normally distributed with mean zero and precisions τ_{0S} and τ_{0R}, respectively. The additional covariates are

1. A dummy indicating if the country was a member of the Schengen agreement as of 1 January 2007, NSn_i.
2. Share of stocks of migrants born outside the EU and the EFTA countries, MS_{0i}/P_i and MS_{0j}/P_j. Source: Parsons et al. (2005).
3. Share of the population older than 65 years, $P65_{it}$. Source: Population Reference Bureau's World Population Data Sheet 2002-2008[5].
4. Life expectancy at birth of women in years. Source: Population Reference Bureau's World Population Data Sheet 2002-2008[6].

The joint density of the flows and parameters, given the covariates, X, in the migration model is

$$\begin{aligned}f(y, z, \mu, \delta, \lambda, \kappa, \varsigma, \tau^k, \tau_m, \tau_y, \tau_{0k}, \alpha, \beta) = f(z|\mu) f(\mu|y, \lambda, \delta, \kappa, \varsigma, \tau^k, \tau_m) \\ \times f(y|\alpha, \beta, \tau_y, \tau_{0k}; X)\ f(\lambda) f(\delta) f(\kappa) f(\tau^k) f(\alpha) f(\beta) f(\tau_{0k}) f(\tau_y) f(\varsigma|\tau_m) f(\tau_m).\end{aligned} \tag{14}$$

where $f(z|\mu)$ is the data model, $f(\mu|y, \lambda, \delta, \kappa, \varsigma, \tau^k, \tau_m)$ is the measurement model, $f(y|\alpha, \beta, \tau_y, \tau_{0k}; X)$ is the migration model and $f(\lambda)$, $f(\delta)$, $f(\kappa)$, $f(\tau^k)$, $f(\alpha)$, $f(\beta)$, $f(\tau_{0k})$, $f(\tau_y)$, $f(\varsigma|\tau_m)$ and $f(\tau_m)$ are the priors. The distribution of y given the observed flows, z, can be obtained by integrating every other parameter out of density given in equation (14).

[3] http://www.llv.li/amtsstellen/llv-as-aussenhandel/llv-as-aussenhandel-direktimporte_nach_laender.htm, accessed July 2010

[4] http://www.liechtenstein.li/en/pdf-fl-multimedia-information-industriestandort-druck.pdf, accessed July 2010

[5] http://www.prb.org, accessed February 2010

[6] http://www.prb.org, accessed February 2010

4 Data Collection

The data used in the project comes primarily from the Eurostat migration data base. These migration data are collected by means of an annual questionnaire (i.e., the Joint Questionnaire on Migration Statistics), which is sent to all national statistical agencies in the European Union. This questionnaire is coordinated by the Council of Europe, the UN Statistical Division, the UN Economic Commission for Europe and the International Labour Organization. Apart from the EU countries, data are also collected for various other European countries, such as Iceland, Norway, Switzerland and Turkey. The variables include age, sex, country of previous or next residence and country of citizenship. Additional data may be obtained from websites organised and maintained by national statistical agencies. Of particular importance to this project is a recent publication on European migration by Poulain et al. (2006), which describes the current situation and sources of international migration data in Europe in great detail.

For our model, we use data on emigration and immigration flows amongst the 31 countries in the EU and EFTA, as well as flows to and from rest of world, from 2002 to 2008. The following assumptions with respect to the data have been made:

- For the Netherlands - category 'Unknown' in the data on emigration was distributed proportionally to all the countries.
- Category 'ex-Czechoslovakia' in the emigration from and immigration to Denmark was distributed to the Czech Republic and Slovakia proportionally in a given year.

5 Constructing the priors

In this project, research is undertaken to design a realistic and effective migration model as described above, where available expert opinion can be conveniently incorporated and estimates and measures of precision efficiently computed. While the proposed extensions provide more realistic and flexible models for migration patterns, this comes at a price: the additional parameters required may be weakly identified from the data. However, the Bayesian approach permits expert opinion to be combined with the data to strengthen the inference. The Bayesian approach also facilitates the combination of multiple data sources, with their differing levels of error, as well as prior information about the structures of the migration processes, into a single prediction with an associated measure of uncertainty.

For the measurement model the priors for duration of stay, undercount and precision are elicited from the experts by means of a Delphi survey. As the elicitation process is in progress, all expert-based prior densities and results presented here should be treated as preliminary only.

For the duration of stay parameters δ_1, δ_2 and δ_3 we assume a mixture of log-normal priors for auxiliary parameters d_k, which results from experts answers. The resulting interquartile ranges and medians for the mixtures are presented in Table 1. The interpretation of these priors is straightforward. Let us take the six months duration criterion, i.e. δ_3 median, 1.22. It means that in median experts expect flows measured using this criterion to be larger than the true flows by 22% with a 12 months criterion definition (or the true flows to constitute 82% of the observed flows). Prior for δ_4 is not elicited from the experts and for its auxiliary parameter d_4 the prior was

$$d_4 \sim \log \mathcal{N}(-0.5, 30),$$

which for δ_4 implies the interquartile range $(0.50,\ 0.59)$ and median of 0.55.

The priors for the undercount parameters λ_r are mixtures of beta densities reflecting experts' opinions about undercount. Their characteristics are presented in Table 2.

The prior densities for precision of the error term in the measurement equations are constructed based on experts opinions as mixtures of gamma densities. Due to the heterogeneity of expert's judgements the resulting priors are rather vague, the interquartile

Table 1: Characteristics of the expert-based priors for duration of stay parameters

	$q(0.25)$	median	$q(0.75)$
δ_1	1.86	2.32	3.49
δ_2	1.41	1.70	2.52
δ_3	1.12	1.22	1.46

Table 2: Characteristics of the expert-based priors for undercount parameters

	$q(0.25)$	median	$q(0.75)$
λ_1& λ_3	0.35	0.58	0.78
λ_2& λ_4	0.64	0.80	0.87

ranges are (14, 982) for emigration and (210, 1373) for immigration equation, with medians 551 and 877, respectively. The median results can be interpreted in the vein of Bijak and Wiśniowski (2010) as allowing deviations of the flows from their average level of ±4.3% and ±3.4%, respectively.

For the random effects (coverage) parameters, κ_i, we assume the following

$$\kappa_{i \text{ is poor}} \sim \mathcal{N}(m.k_1, t.k_1),$$

$$\kappa_{i \text{ is good}} \sim \mathcal{N}(m.k_2, t.k_2),$$

where

$$m.k_1 \sim \mathcal{N}(1, 0.5),$$
$$m.k_2 \sim \mathcal{N}(2, 0.5),$$

and

$$t.k_1 \sim \mathcal{G}(4, 1),$$
$$t.k_2 \sim \mathcal{G}(4, 4).$$

This specification is consistent with the fifth approach for coverage as specified in the Appendix. For countries with *excellent* coverage (DK, FI, NL, NO, SE) we assume random effects are fixed and always equal to zero on the log scale (hence the resulting scaling factor is always 1). These priors imply the coverage random effects characteristics to be as in Table 3. This subjective specification is based on our experiences in the MIMOSA project but in fact is rather weakly informative.

Table 3: Characteristics of the priors for random parameters

	$q(0.25)$	median	$q(0.75)$
poor coverage	0.50	0.73	0.88
good coverage	0.69	0.88	0.96

For the constants in the migration models for intra-European flows a normal hierarchical prior is assumed

$$\alpha_1 \sim \mathcal{N}(0, \tau_\alpha),$$
$$\tau_\alpha = 1/a^2, \quad a \sim \mathcal{U}(1, 10).$$

Same priors for constants in the model for flows between EU/EFTA and rest of world is taken. This prior reduces the autocorrelation of the MCMC sample greatly and allows for faster convergence of the algorithm. For the rest of the parameters in the migration models weakly informative normal priors, $\mathcal{N}(0, 0.1)$, are assumed. For the precision in the migration models we assume a gamma prior density $\mathcal{G}(1, 1)$.

6 Preliminary results

The model has been tested using OpenBUGS software dedicated for Bayesian computations. The posterior characteristics were computed basing on the MCMC samples of 10,000 length with 5,000 burn-in sample.

To assess how the model predicts data on emigration and immigration, we compared the means of the posterior distributions for μ^S and μ^R with the available data. Table 4 presents the absolute, mean absolute and root mean square errors for all 31 countries and seven years. The sums of flows are given for comparison. We observe that the model predicts the data well. The relative absolute error is 0.05-0.06%. The RMSE shows there may be some larger deviations from the data but they concern flows of tens or hundreds of thousand people.

Table 4: Prediction of the data by the model

data	AE	MAE	RMSE	Sum of flows
Immigration	3786	0.91	1.98	7,870,000
Emigration	3363	0.79	1.58	5,460,000

As far as the parameters of the model are concerned, we present an example posterior characteristics of the duration of stay factors. In Table 5 the posterior means of the duration of stay factors are presented. These factors are computed by exponentiating the negative of a posterior MCMC sample of the parameters δ. Hence, we can interpret them as a factor in the equation $true\ flow = factor \times data$. Our benchmark criterion of 12 months produces a factor equal to one. For countries with a 'no time limit' of stay criterion, the true flows constitute 48% of the observed data. For six months, this factor is 72%. For permanent the true flows are on average 1.78 times larger than the observed data. The posterior densities of the duration of stay factors are presented in Figure 1.

Table 5: Posterior characteristics of the duration criteria factors

duration	no time limit	3 months	6 months	**12 months**	permanent
mean	48%	60%	72%	**100**%	178%
std.dev	4.0%	3.9%	3.2%	NA	15.0%

The undercount of emigrants with respect to immigrants is on average 59% (standard deviation was 4.2%) for emigration and 84.5% (with standard deviation 5.6%) for immigration within EU/EFTA countries. We can interpret that as the observed emigration data constitute 59% of the true flows and the observed immigration data account for 84.5% of the true flows. The undercount of flows to the rest of the world equals a posteriori 62.6% (with standard deviation 14.2%) and of flows from the rest of world 82.4% (with standard deviation 7.7%). As the identification of undercount parameters for emigration and immigration from the data exclusively is impossible, the expert-based priors were the main source of information about these parameters in our model.

The ultimate goal of our analysis is a table of the posterior distributions of all the true flows between 31 countries within a period 2002-2008. In Table 6 and Figure 2 we present posterior characteristics and densities of 2007 flows between Denmark (DK) and The Netherlands (NL) and flows between France (FR) and Hungary (HU).

Table 6: Posterior characteristics of the migration true flows

flow	mean	std.dev	median
DK-NL	443	64	438
FR-HU	1530	1382	1116

Figure 1: Posterior densities of the duration criteria parameters

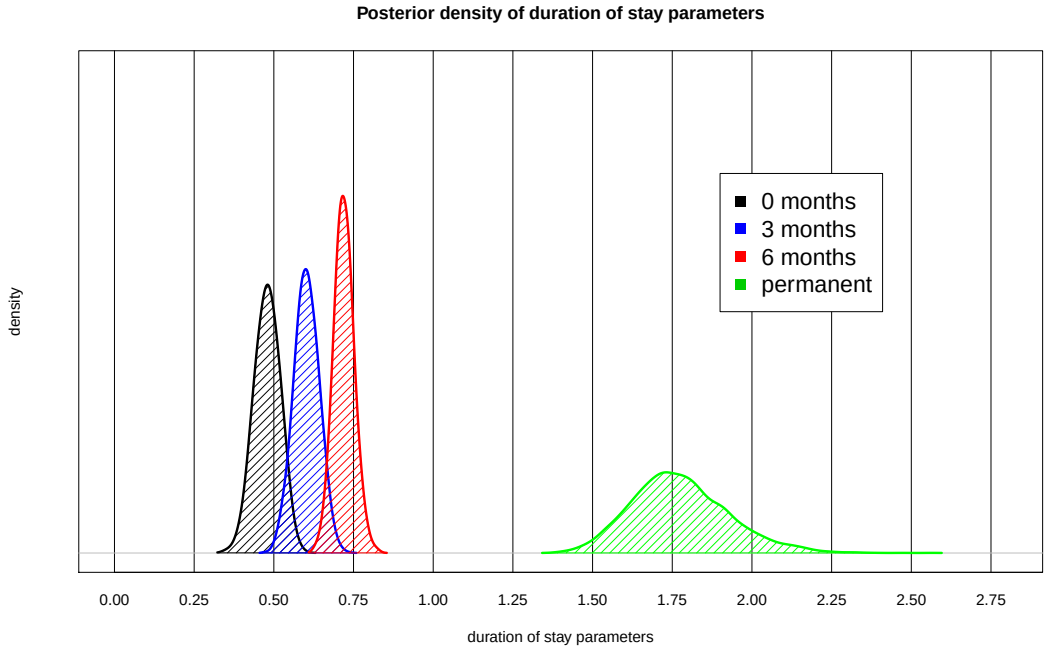
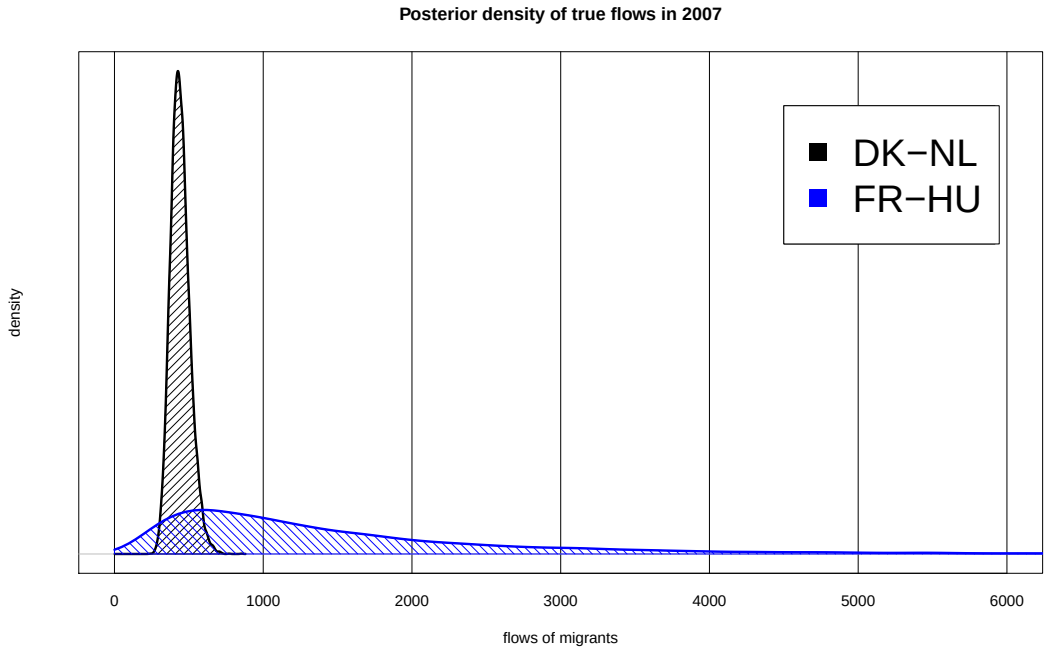


Figure 2: Posterior densities of the migration true flows



For flows between DK and NL both data on immigration and emigration are available, the resulting prior is comparatively tight, with mean around 440 people. The posterior density of the flow between FR and HU is based mainly on the information from the migration model, as no data has been reported. This flow is characterised by a heavy tail. This may be a result of only a limited information carried by the covariates as well as their inner variability.

7 Conclusions

This paper has presented the initial model framework of the IMEM project. Prototype testing has been done on the whole set of countries and all years envisaged and it appears to be a promising approach. The models and MCMC algorithms are being programmed in both R and OpenBUGS. The next steps for the project are to continue developing the model and to further elicit expert information on the definition, coverage and accuracy aspects of the model to improve the specification of the priors.

References

[1] Abel GJ (2010) Estimation of international migration flow tables in Europe. Journal of the Royal Statistical Society, Series A, 173.

[2] Bijak J, Wiśniowski A (2010) Bayesian forecasting of immigration to selected European countries by using expert knowledge. Journal of the Royal Statistical Society, Series A, 173.

[3] Brierley MJ, Forster JJ, McDonald JW and Smith PWF (2008) Bayesian estimation of migration flows. In *International Migration in Europe: Data, Models and Estimates*, pp. 149-174. Wiley: Chichester.

[4] de Beer J, Raymer J, van der Erf R and van Wissen L (2010) Overcoming the Problems of Inconsistent International Migration data: A New Method Applied to Flows in Europe. European Journal of Population 26:459-481.

[5] Congdon P (2001) The development of gravity models for hospital patient flows under system change: a Bayesian modelling approach. Health Care Management Science 4(4):289-304.

[6] Council of Europe (2002) Recent demographic developments in Europe. Strasbourg: Council of Europe.

[7] Czado C, Delwarde A and Denuit M (2005) Bayesian Poisson log-bilinear mortality projections. Insurance: Mathematics and Economics 36(3):260-284.

[8] Daponte BO, Kadane JB and Wolfson LJ (1997) Bayesian Demography: Projecting the Iraqi Kurdish Population, 1977-1990. Journal of the American Statistical Association 92:1256-1267.

[9] van der Erf R (2009) Typology of Data and Feasibility study, MIMOSA Deliverable 9.1 B Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

[10] Girosi F and King G (2008) Demographic Forecasting. Princeton: Princeton University Press.

[11] Gorbey S, James D and Poot J (1999) Population Forecasting with Endogenous Migration: An Application to Trans-Tasman Migration. International Regional Science Review, 22(1):69-101.

[12] Hill K (1985) Indirect approaches to assessing stocks and flows of migrants. *In Immigration statistics: A story of neglect*, Levine DB, Hill K and Warren R, eds., pp. 205-224. Washington, DC: National Academy Press.

[13] Jasso G and Rosenzweig MR (1982) Estimating the emigration rates of legal immigrants using administrative and survey data: The 1971 cohort of immigrants to the United States. Demography 19:279-290.

[14] Jennissen R (2004) Macro-economic determinants of international migration in Europe. PhD Thesis, Rijksuniversiteit Groningen.

[15] Kupiszewska D and Wiśniowski A (2009) Availability of statistical data on migration and migrant population and potential supplementary sources for data estimation, MIMOSA Deliverable 9.1 A Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

[16] Lynch SM (2007) Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer.

[17] Mayer T and S Zignago (2006) Notes on CEPIIs distances measures. Centre dÉtudes Prospectives dÍnformations Internationales (CEPII), Paris.

[18] Nowok B, Kupiszewska D and Poulain M (2006) Statistics on international migration flows. In *THESIM: Towards harmonised European statistics on international migration*, Poulain M, Perrin N and Singleton A, eds., pp. 203-231. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

[19] Parsons CR, R Skeldon, TL Walmsley, and LA Winters (2005) Quantifying the International Bilateral Movements of Migrants. 8th Annual Conference on Global Economic Analysis, Lübeck, Germany, June 9-11

[20] Poulain M (1993) Confrontation des statistiques de migration intra-européennes: vers une matrice complète? European Journal of Population 9(4):353-381.

[21] Poulain M (1999) International migration within Europe: Towards more complete and reliable data? Working Paper No. 37, Conference of European Statisticians, Statistical Office of the European Communities (Eurostat), Perugia, Italy.

[22] Poulain M, Perrin N and Singleton A, eds. (2006) THESIM: Towards Harmonised European Statistics on International Migration. Louvain: UCL Presses.

[23] Raymer J (2007) The estimation of international migration flows: A general technique focused on the origin-destination association structure. Environment and Planning A 12:371-388.

[24] Raymer J (2008) Obtaining an overall picture of population movement in the European Union. In International migration in Europe: Data, models and estimates, Raymer J and Willekens F, eds., pp. 209-234. Chichester: Wiley.

[25] Raymer J, de Beer J, van der Erf R (forthcoming) Putting the Pieces of the Puzzle Together: Age and Sex-Specific Estimates of Migration amongst Countries in the EU/EFTA. European Journal of Population.

[26] Schmertmann CP (1992) Estimation of historical migration rates from a single census: Interregional migration in Brazil 1900-1980. Population Studies 46(1):103-120.

[27] Tuljapurkar S and Boe C (1999) Validation, probability-weighted priors and information in stochastic forecasts. International Journal of Forecasting 15(3):259-271.

[28] United Nations (1998) Recommendations on statistics of international migration. Statistical Papers Series M, No. 58, Rev.1, Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

[29] Van der Gaag N and Van Wissen L (2002) Modelling regional immigration: Using stocks to predict flows. European Journal of Population 18:387-409.

[30] Warren R and Peck JM (1980) Foreign-Born Emigration from the United States: 1960 to 1970. Demography 17(1):71-84.

[31] Willekens F (1994) Monitoring international migration flows in Europe. Towards a statistical data base combining data from different sources. European Journal of Population 10(1):1-42.

[32] World Bank (2010) World Development Indicators. Accessed at http://data.worldbank.org/data-catalog/world-development-indicators, The World Bank, Washington.

[33] Zaba B (1987) The Indirect Estimation of Migration: A Critical Review. International Migration Review 21(4):1395-1445.

Appendix: Coverage issues

During the testing of the model various versions of introducing random effects or coverage to the model have been considered.

In the first approach we assumed that the coverage coefficient for a given country is a product of an unconstrained country specific random effect normally distributed around zero ($re_m = \exp(\varsigma_i)$ on a linear scale) and a group-specific mean coverage ($kp_m = (1 + e^{-\kappa_m})^{-1}$), constrained to be in the range $(0, 1)$. Three groups were assumed: poor, good and excellent. The precisions for random effects are grouped in order to reduce the parameter space. The excellent group, containing four Scandinavian countries (DK, FI, NO, SE) and the Netherlands (NL), has a prior with a large precision. For the two other groups, the priors are rather vague. In order to ensure identification of the parameters, the coverage mean of the excellent group is treated as a baseline, i.e. poor and good groups are measured relatively to the excellent one. Then, for poor and good countries mean coverage is a product of the form $kp_{\text{poor}} \times kp_{\text{excellent}}$ and $kp_{\text{good}} \times kp_{\text{excellent}}$, respectively.

In the second approach the mean coverage for the excellent group is set to 1, the rest of the parameterisation remains as described above. The third option allows the coverage effects to be captured only by unconstrained random effects ($\exp(\varsigma_i)$). The fourth approach uses a fixed country-specific coverage, that is $(1+e^{-\kappa_i})^{-1}$, without random effects. The fifth approach introduces the logistic-normal random effects. We allow the underlying normal density of the country-specific coverage to have a free mean and a free precision. In order to preserve parsimonious parametrisation, the precisions of these effects are grouped as in the first option. The current model specification, described in Section 3.1, is even more parsimonious by grouping the means of the coverage effects. The last approach introduces an additional country-specific random effect which, for a given flow, measures the effect of the country not taking part in counting of migrants. For simplicity no constrained coverage is assumed. Thus in the measurement equations we have

$$\log \mu^S_{ijt} = \log y_{ijt} + \beta_i + \varsigma_i + \phi_j + \varepsilon^S_{ijt}, \quad \forall_t \forall_{i,j} \tag{15}$$

$$\log \mu^R_{ijt} = \log y_{ijt} + \gamma_j + \phi_i + \varsigma_j + \varepsilon^R_{ijt}, \quad \forall_t \forall_{i,j}, \tag{16}$$

where ς_i is a 'usual' measuring-country-specific random effect and ϕ_j is a non-measuring-country-specific random effect. They are normally distributed $\mathcal{N}(0, t_\varsigma)$ and $\mathcal{N}(0, t_\phi)$, respectively.

Priors used in the analysis are described below.

1. Prior densities assumed for coverage auxiliary parameter κ are:

$$\kappa_{\text{poor}} \sim \mathcal{N}(1, 4),$$

$$\kappa_{\text{good}} \sim \mathcal{N}(2, 1),$$

$$\kappa_{\text{excellent}} \sim \mathcal{N}(5, 0.25).$$

Priors for the precisions of the random effects:

$$\tau_{\text{poor}} \sim \mathcal{G}(1,1),$$

$$\tau_{\text{good}} \sim \mathcal{G}(2,0.1),$$

$$\tau_{\text{excellent}} \sim \mathcal{G}(10000,5),$$

which results in relatively vague priors for countries suspected to poorly cover persons who should be migrants. For countries where it is believed that the coverage is nearly perfect, the precision is set to be comparatively high. In terms of a relative standard error (assuming log-normal distribution of μ), it can be explained as a relative standard deviation of the observed migration from the level of the true flow, following Bijak and Wiśniowski (2010) approach. The priors imply $\pm 130\%$ for poor, $\pm 22\%$ for good and $\pm 2\%$ for excellent types.

2. Priors for the mean coverage in the groups *poor* and *good* and the priors for the precisions of the random effects are the same as in 1. The coverage for the *excellent* group is set to one.

3. Priors for the random effects are the same as in 1. There is no constrained coverage in this approach.

4. Priors for coverage are the same as in 1. There are no random effects.

5. Random effects are logistic-normal, thus κ_m has a normal density with country-specific mean and group-specific precision, i.e.

$$\mathcal{N}(k_i, t.k_m).$$

The priors for k_i are

$$k_{i \text{ is poor}} \sim \mathcal{N}(1, t.k_{\text{poor}}),$$

$$k_{i \text{ is good}} \sim \mathcal{N}(2, t.k_{\text{good}}),$$

$$k_{i \text{ is excellent}} \sim \mathcal{N}(5, t.k_{\text{excellent}}).$$

For precision, the priors are

$$t.k_{\text{poor}} \sim \mathcal{G}(4,1),$$

$$t.k_{\text{good}} \sim \mathcal{G}(4,4),$$

$$t.k_{\text{excellent}} \sim \mathcal{G}(4,4).$$

6. There is no constrained coverage. Counting-country random effects ς_i are specified same as in 1. non-measuring-country random effects ϕ_j have zero mean and the prior for their precision is

$$\tau_\phi \sim \mathcal{G}(1,1).$$

Figure 3 presents the posterior results of all the model versions. On the X axis there are three groups: poor, good and excellent. On the Y axis the median of the overall effect of coverage on the true flow is presented. Median was chosen as a robust point estimator of coverage effect, as the resulting posterior densities, especially in the cases with unconstrained random effects for the countries with missing data, had heavy tails. For the first and second cases it is the product of the random effect and the mean group-specific coverage, for third to sixth it is median of a random effect.

The general pattern for all countries is observed in all figures. Countries with assumed excellent coverage are centered around 1. There is hardly any difference between the first

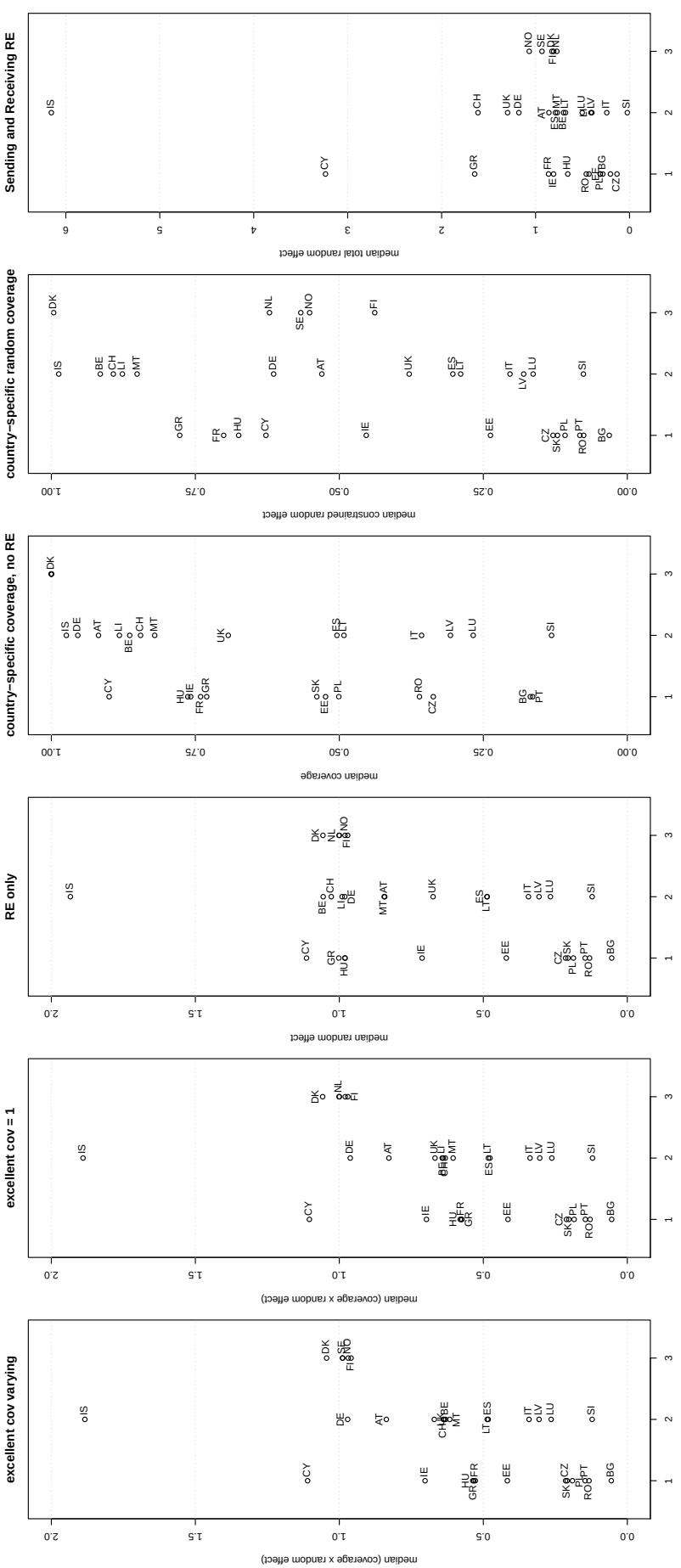


Figure 3: Coverage effect in five parameterisations

and second approach. The only effect was some reduction in the correlation of the random effects and mean coverage chains.

The third parametrisation, with free random effects only, is surprisingly close to the first two ones. The only difference concerns the countries for which no data at all are available (BE, CH, HU, GR, FR, LI, MT). Their effects are centred around 1, instead of the mean coverage implied by the prior density in the first two approaches. For the countries with the data available there is almost no difference in posterior (unconstrained) coverage effect.

In the fourth parameterisation the pattern for the countries with the data available is again similar, their coverage is close to where it used to be in first three approaches. For the countries with no data, the coverage is around the values implied by the prior densities.

Fifth approach turns out to deviate mostly from the pattern. Countries with assumed perfect coverage are way below 100%, only DK coverage remains excellent. In the 'good' category, for missing data countries: CH, MT, LI, BE, coverage looks surprisingly well (their medians are around the prior median for random coverage), while countries such as DE or AT have lower posterior coverage (around 50-60%). The same situation is in 'poor' coverage group, FR, HU and GR are centered around the expected mean of a random coverage as in the prior. On the other hand, a significant increase can be noticed in true flows posetrior estimates. They are on average larger in the case of random coverage than in any of the former cases, even with fixed coverage. It should be noted that the prior for the excellent group can be tighter which may potentially allow to shift the excellent group close to one.

Suspicious results are obtained for CY and IS, where median coverages are above one in all unconstrained coverage effects approaches. In the fifth approach IS coverage remains close to 100% but for CY it falls to about 60%.

In the sixth approach (double random effects), the excellent group is characterised by larger spread. IS and CY are, respectively, a posteriori six and three times overcounted in the reported data, comparing to the true flows. Moreover, GR, DE, UK and NO seem to have overcount in their data. Summarising, this approach, at least unconstrained, seems to be inappropriate for capturing heterogeneity across countries, interpreted in terms of coverage.

Summarising, allowing the unconstrained random effects in the model may lead to too optimistic posterior estimate of the undercount resulting from coverage and other country-specific effects, having controlled for emigration undercount (λ) and duration of stay criteria (δ). It should be noted, that the other measurement equation parameters remain stable in almost all versions of coverage.