Integrated Modelling of European Migration

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

Southampton Statistical Sciences Research Institute
University of Southampton

October 14, 2011

— DRAFT —
Note: we are currently extending this paper to include age and sex structures

Abstract

International migration data are collected by individual countries with separate collection systems and designs. As a result, reported data are inconsistent in availability, definition and quality. Rather than wait for countries to harmonise their migration data collection and reporting systems, we propose a model to overcome the limitations of the various data sources. In particular, 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) from 2002 to 2008, using data collected by Eurostat and other national and international institutions. We also incorporate additional information provided by experts on the effects of undercount, measurement and accuracy of data collection systems. The methodology is integrated and produces a synthetic data base with measures of uncertainty for international migration flows and other model parameters.

1 Introduction

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 the availability, definitions and quality (Kelly 1987; Zlotnik 1987; Willekens 1994; Bilsborrow et al. 1997; Poulain et al. 2006; Kupiszewska and Nowok 2008). 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) from 2002 to 2008, using data collected by

*Contact email at raymer@soton.ac.uk

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 levels of migration. 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, our 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.

This paper is structured as follows. First, we provide some background and context to this work. Second, in Section 3, we specify the model, which includes a model for measurement error and a spatial interaction-type model for estimating missing migration flows. In Section 4, we describe the main sources of the data used in this paper. The construction of the priors is then presented in Section 5, followed a presentation of the results. Finally, the paper concludes with a summary and a discussion.

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 no minimum period, three months, six months, one year, or even more. Recently, there have been several papers addressing this issue (see, e.g., de Beer et al. 2010; Nowok 2010; Nowok and Willekens 2011).

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

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

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 (Raymer and Bijak 2009; Raymer et al. 2011). 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 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., Warren and Peck 1980; Jasso and Rosenzweig 1982; Hill 1985; Zaba 1987; van der Gaag and van Wissen 2002). There are, however, several (mostly recent) efforts on estimating migration flow matrices from which we can draw experiences. First, Poulain (1993, 1999) developed a correction factor approach, using optimisation techniques, to reconcile differences between flows reported by receiving and sending countries in Europe. The calculated correction factors provided both a means for harmonising migration flows and a basis for understanding the reasons for the differences. Second, Raymer (2007, 2008) developed a hierarchical multiplicative component approach for estimating international migration flows in Europe by age and sex. The multiplicative component approach showed how log-linear models could be applied to model international migration flows in a systematic and hierarchical manner. Third, Brierley et al. (2008) extended this approach in a Bayesian context and demonstrated the usefulness and flexibility of incorporating various forms of prior information and the importance of distributions quantifying uncertainty in the estimated values. Fourth, Cohen et al. (2008) developed a gravity model to project international migration flows for all countries in the world (see also Kim and Cohen 2010). Fifth, Abel (2010) extended Poulain's optimisation methods and applied statistical methods for missing data (i.e., expectation-maximisation) to estimate flows amongst 15 European Union countries.

Finally, and most recently, researchers at the Netherlands Interdisciplinary Demographic Institute (NIDI, The Hague), the Central European Forum for Migration and Population Research (CEFMR, Warsaw), the Southampton Statistical Sciences Research Institute (S3RI) and the Université Catholique de Louvain (Charleroi) collaborated on a Eurostat-funded project to estimate international migration stocks and flows in Europe. The methodology adopted by the MIMOSA (MIgration MOdelling for Statistical Analyses) team represented a two-stage hierarchical procedure. The first stage, described in de Beer et al. (2010), harmonises the flows amongst 19 EU/EFTA countries providing both sending and receiving data by applying an extension of Poulain's (1999; see also Abel 2010) optimisation procedure benchmarked to Sweden's migration flow data, which were assumed to be measured more or less without error. The second stage, described in Raymer et al. (2011), estimates the missing marginal data and associations between countries by using the harmonised 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 the hierarchical estimation procedure. However, the estimates produced are considered reasonable and currently

represent the most extensive and comprehensive set of harmonised migration flow estimates in Europe. These estimates may be downloaded from NIDI's website. [2]

The above research has led us to the conclusion that a Bayesian approach offers the best opportunity for integrating all the different types of data, covariate information 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, as noted by Willekens (1994), the methodology provides a formal mechanism for the inclusion of expert judgement to supplement the deficient migration data.

Applications of Bayesian methods in migration and population forecasting include predictions of international migration from time series models (Gorbey et al. 1999; Bijak and Wiśniowski 2010; Bijak 2011; Mitchell et al. 2011). They have also been used to model non-migratory spatial movements (Congdon 2001), produce forecasts of fertility (Tuljapurkar and Boe 1999) and mortality (Czado et al. 2005; Girosi and King 2008; Chunn et al. 2010), and to estimate population sizes under situations of very limited information (Daponte et al. 1999). 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. We observe counts (flows) z^k_{ijt} from country i to country j during year t reported by either the sending S or receiving R country, where $k \in \{S, R\}$. These flows can be represented by matrices Z^S_t and Z^R_t:

$$Z^S_t = \begin{pmatrix} 0 & z^S_{12t} & z^S_{13t} & \cdots & z^S_{1nt} \\ z^S_{21t} & 0 & z^S_{23t} & \cdots & z^S_{2nt} \\ z^S_{31t} & z^S_{32t} & 0 & \cdots & z^S_{3nt} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z^S_{n1t} & z^S_{n2t} & z^S_{n3t} & \cdots & 0 \end{pmatrix} \quad Z^R_t = \begin{pmatrix} 0 & z^R_{12t} & z^R_{13t} & \cdots & z^R_{1nt} \\ z^R_{21t} & 0 & z^R_{23t} & \cdots & z^R_{2nt} \\ z^R_{31t} & z^R_{32t} & 0 & \cdots & z^R_{3nt} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z^R_{n1t} & z^R_{n2t} & z^R_{n3t} & \cdots & 0 \end{pmatrix}.$$

The interest of this research is to estimate a matrix Y_t of true migration flows with unknown entries:

$$Y_t = \begin{pmatrix} 0 & y_{12t} & y_{13t} & \cdots & y_{1nt} \\ y_{21t} & 0 & y_{23t} & \cdots & y_{2nt} \\ y_{31t} & y_{32t} & 0 & \cdots & y_{3nt} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1t} & y_{n2t} & y_{n3t} & \cdots & 0 \end{pmatrix}.$$

[2] http://www.nidi.nl/Pages/NID/24/928.bGFuZz1VSw.html.

For all i, j and t, we assume that z_{ijt}^k follows a Poisson distribution

$$z_{ijt}^S \sim \text{Po}(\mu_{ijt}^S), \quad (1)$$
$$z_{ijt}^R \sim \text{Po}(\mu_{ijt}^R). \quad (2)$$

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.*

The two measurement error equations are

$$\log \mu_{ijt}^S = \log y_{ijt} + \psi_i - \log\left(1 + e^{-\kappa_i}\right) + \varepsilon_{ijt}^S, \quad (3)$$
$$\log \mu_{ijt}^R = \log y_{ijt} + \gamma_j - \log\left(1 + e^{-\kappa_j}\right) + \varepsilon_{ijt}^R, \quad (4)$$

where we assume $\varepsilon_{ijt}^S \sim \mathcal{N}(0, \tau_i^S)$ and $\varepsilon_{ijt}^R \sim \mathcal{N}(0, \tau_j^R)$. The precisions (reciprocal variances) of the error terms depend on whether the data are captured by sending or receiving countries. Thus we take

$$\tau_i^S = t_{c(i)}^S, \quad (5)$$
$$\tau_j^R = t_{c(j)}^R, \quad (6)$$

where $c(i)$ denotes the type of collection system (e.g., population register or survey). The number of parameters required to capture differences in accuracy depends on our typology of collection systems, and their relative ability to capture migration flows, regardless of definition and coverage. We distinguish three types of systems: 1) registers in the the Scandinavian countries (DK, FI, IS, NO, SE), which are exchanging the data about migrants among themselves, 2) the register based systems in the other countries and 3) survey based systems in the UK, CY and IE. For all three types no constraints are assumed.

The differences in duration of stay criterion, which depend on the reporting country, and the effect of undercount are captured by the parameters ψ_i and γ_j,

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

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

The δ_m parameter measures the effect of a particular duration of stay definition used by country i. The parameters are constrained so that $\delta_1 > \delta_2 > \delta_3 > 0$ and $\delta_4 < 0$

in the following way,

$$\begin{aligned}\delta_1 &= d_1 + d_2 + d_3,\\ \delta_2 &= d_2 + d_3,\\ \delta_3 &= d_3,\\ \delta_4 &= -d_4,\end{aligned}$$

where $d_k > 0$ are auxiliary parameters. The λ_r parameters measure the effect of the undercount with the assumption that $\lambda_r \in (0, 1)$.

In order to handle the huge differences between immigration and emigration data reported by ES and IT, an unconstrained dummy is included in the equation for both intra-EU and rest of world emigration measurement. For Spain, the reason for treating its definition separately lies in the data collection methods (González-Ferrer, 2009). Emigrants are counted since 2002 through self-reporting in the country of destination. Since 2004, the statistics include also those who have been deregistered ex officio by the municipalities. This shift is observed in the category Unknown, which in 2002 was 6 persons, in 2003 38,000, reaching in 2008 almost 200,000 emigrants. Note, that the numbers of immigrants are also deemed to be heavily underestimated (González-Ferrer, 2009). In Italy, according to the PROMINSTAT report (Gabrielli et al. 2009), the measurement of emigration is based on deregistration and corrections (so called cancellations). There is no duration of stay criterion but it is considered that the measurement concerns the long-term migrations according to the UN definition. However, the cancellations only refer to the permanent movements, which may lead to a heavy undercount of emigrants.

Finally, the κ_i parameter is a normally distributed country-specific random effect

$$\kappa_i \sim \mathcal{N}(\nu_i, \zeta_i),$$

where $\nu_i = \nu_{m(i)}$ is a group-specific mean, $\zeta_i = \zeta_{m(i)}$ is a group-specific precision and $m(i)$ denotes a type of coverage assumed for country i. For the time being, there are two coverage types, that is, $m(i) \in \{\text{standard}, \text{excellent}\}$. The logistic transformation of κ in Equations 3 and 4 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 migration to and from the rest of world there is only one equation per outflow and inflow, respectively, i.e.,

$$\log \mu^S_{i0t} = \log y_{i0t} + \psi_i + \varepsilon^S_{i0t}, \quad \text{for all } i \text{ and } t \tag{9}$$

$$\log \mu^R_{0jt} = \log y_{0jt} + \gamma_j + \varepsilon^R_{0jt}, \quad \text{for all } j \text{ and } t, \tag{10}$$

All other parameters remain same as described above, except for ψ_i and γ_j, which are defined as in Equations 7 and 8 with λ_1 and λ_2 replaced with λ_3 and λ_4, respectively. Note, that in the measurement of the flows to and from the rest of world we assume a perfect coverage for all countries, i.e., there are no country-specific random effects.

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, e.g., Jennissen 2004; Abel 2010; Raymer et al. 2011).

The explanatory variables can be grouped into economic, demographic and geographic ones. Consider the following model of migration:

$$\begin{aligned}\log y_{ijt} = \alpha_1 &+ \alpha_2 \log P_{it} + \alpha_3 \log P_{jt} + \alpha_4 C_{ij} + \alpha_5 \log T_{ijt} + \alpha_6 \log(G_{it}/G_{jt}) \\ &+ \alpha_7 A_{ijt} + \alpha_8 A_{it} + \alpha_9 A_{jt} + \alpha_{10} S_{ij} + U_{ij} \\ &+ \alpha_{11} E_2 + \alpha_{12} E_3 + \alpha_{13} E_4 + \alpha_{14} E_5 + \alpha_{15} E_6 + \alpha_{16} E_7 + \xi_{ijt}, \end{aligned} \tag{11}$$

where $\alpha = (\alpha_1, \ldots, \alpha_{16})'$ 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 are assumed.

3. The ratio of the Gross National Income per capita in sending and receiving countries, G_{it}/G_{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 [3].

5. Three dummies for accession, meant to capture the changing EU/EFTA membership status between 2002 and 2008. The first one, A_{ijt}, takes the value 1 if both i and j were in the EU/EFTA in year t. The second one, A_{it}, is 1 if a sending country was in the EU/EFTA in year t. The third, A_{jt}, is 1 if a receiving country was in the EU/EFTA in year t.

6. Origin-destination migrant stocks based on the 2000 population censuses round, S_{ij}; source: Parsons et al. (2005).

7. Year dummies, E_t, $t = 2, \ldots, 7$ are introduced in order to handle the different levels of migration for different years, 2002 to 2007. The reference year is 2008.

8. In order to smooth the data over time, flow specific but constant over time random effects are introduced. They are denoted as U_{ij} and are normally distributed with mean zero and a common precision, that is $U_{ij} \sim \mathcal{N}(0, \tau_u)$. With the year dummies E_t present, they capture the individual flow pattern and do not allow for large differences between the data and predicted values, especially when the particular years in the data are missing.

All non-indicator variables were divided by their means and then logged (migrant stocks are in levels due to zero entries in the data).[4]

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

[4] For Liechtenstein the following imputations were carried out: (1) GNIs per capita were assumed the same as in Switzerland (CH); (2) Trade flows are taken from the statistical office's website, http://www.llv.li, accessed July 2010, and converted from CHF to USD. Exports from Liechtenstein to other countries was approximated. Trade between LI and CH is calculated using ratios as presented in 'Liechtenstein - Industrial location', http://www.liechtenstein.li/en/, accessed July 2010, with the result that exports from LI to CH was estimated to be 11% of the total and the imports 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. (2011).

$$\log y_{i0t} = \beta_1 + \beta_2 \log P_{it} + \beta_3 \log G_{it} + \beta_4 H_i + \beta_5 \log S_{0i} + \beta_6 \log E_{it} + \beta_7 \log L_{it} + U_{ij} + \xi_{i0t}, \quad (12)$$

and for flows from the rest of world

$$\log y_{0jt} = \beta_8 + \beta_9 \log P_{jt} + \beta_{10} \log G_{jt} + \beta_{11} H_j + \beta_{12} \log S_{0j} + \beta_{13} \log E_{jt} + \beta_{14} \log L_{jt} + U_{ij} + \xi_{0jt}. \quad (13)$$

The 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, H_i.

2. Stocks of migrants born outside the EU and the EFTA countries, S_{0i} and S_{0j}; source: Parsons et al. (2005).

3. Share of the population older than 65 years, E_{it}; source: Population Reference Bureau's World Population Data Sheet 2002-2008[5].

4. Life expectancy at birth of women in years, L_{jt}; source: Population Reference Bureau's World Population Data Sheet 2002-2008[6].

5. The flow-specific and constant in time random effects, U_{ij}, are normally distributed with mean zero and precisions τ_{u1} for emigration and τ_{u2} for immigration, respectively. Their purpose, analogously to the intra-EU model, is to smooth the predicted flows across time.

4 Data Collection

For this study, we collected as much data as we could on on the migration flows amongst the 31 countries in the EU and EFTA from 2002 to 2008.[7] Our model includes flows to and from rest of world, which is needed to obtain the total immigration and emigration flows for each country. In the future, we hope to extend the modelling to include age, sex and the year 2009, for which some data are starting to become available (at the time of this writing).

The migration flow data used in the project comes primarily from the Eurostat data base, which relies on the annual Joint Questionnaire on Migration Statistics sent to all national statistical agencies in the European Union. This questionnaire is coordinated by the Council of Europe, the United Nations Statistical Division, the United Nations Economic Commission for Europe and the International Labour Organization. In some cases, we obtained additional information from websites organised and maintained by national statistical agencies. Furthermore, as described

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

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

[7] The following assumptions with respect to the data have been made: (1) For the Netherlands - category 'Unknown' in the data on emigration was distributed proportionally to all the countries; (2) Category 'ex-Czechoslovakia' in the migration from and to Denmark was distributed to the Czech Republic and Slovakia proportionally for each year.

in the previous section, we collected covariate information for use in the migration model to estimate missing flows.

Before constructing the model, we spent a lot of time trying to understand the country-specific measurements underlying the reported statistics. Here, we relied heavily on Poulain et al. (2006), two MIMOSA reports (Kupiszewska and Wiśniowski 2009; van der Erf 2009) and our own analyses (van der Erf 2010). We found that no two countries used the same collection system and often different conceptualisations of migration are utilised. In many cases, these conceptualisations differed within countries, depending on whether the migrant was a foreigner, EU/EFTA member or national of the sending or receiving country.

5 Constructing the priors

In order to produce realistic and meaningful estimates of migration flows, expert opinions are required. The Bayesian approach permits expert opinion to be combined with the data to strengthen the inference. For this study, we sought expert information on undercount, effects of duration and accuracy of migration data collection systems for the purpose of informing our priors for the measurement model. 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.

The priors for duration of stay, undercount and precision are elicited from the experts by means of a Delphi survey (see Wiśniowski et al. 2011). For the duration of stay parameters, δ_1, δ_2, δ_3 and δ_4, we assume a mixture of log-normal priors for auxiliary parameters d_m, which were obtained from the experts. For a more intuitive interpretation we present the characteristics of the terms $\exp(-\delta_m)$. The interquartile ranges and medians for the mixtures are presented in Table 1. We can interpret them as a multiplicative adjustment factor in the equation $true\ flow = factor \times data$. For example, consider the median of six months duration parameter, $\exp(-\delta_3)$, which is equal to 0.81. This value implies that the true flows measured by using a 12 month duration are expected to constitute 81% of the corresponding flows measured with a 6 month duration criterion.

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

	$q(0.25)$	median	$q(0.75)$
$\exp(-\delta_1)$	0.39	0.51	0.60
$\exp(-\delta_2)$	0.50	0.61	0.70
$\exp(-\delta_3)$	0.73	0.81	0.88
$\exp(-\delta_4)$	1.24	1.64	3.55

The priors for the undercount parameters λ_r are mixtures of beta densities resulting from the experts' opinions about undercount (see Wiśniowski et al. 2011). Their characteristics are presented in Table 2.

The prior densities for the precision of the error terms in the measurement equations were also obtained from experts, and then combined into mixtures of gamma densities. Due to the heterogeneity of expert's judgements the resulting priors are rather vague with interquartile ranges of (26, 910) for emigration and (171, 1240) for immigration, and with medians of 573 and 780, respectively. This implies that,

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

	$q(0.25)$	median	$q(0.75)$
λ_1	0.38	0.52	0.68
λ_2	0.60	0.76	0.84
λ_3	0.33	0.54	0.71
λ_4	0.47	0.64	0.77

on average, the flows deviate from their means by $\pm 4.2\%$ and $\pm 3.6\%$, respectively. These same priors were used for all three types of accuracies, thus, the differences between systems in the posterior distributions are based purely on the data.

The coverage random effects parameters, κ_i, for countries with excellent coverage (i.e., DK, FI, NL, NO, SE) are assumed fixed and equal to zero on the log scale. Hence, the resulting scaling factor for the true flows is equal to 1. For the rest of the countries, with standard coverage, we assume the following

$$\kappa_i \sim \mathcal{N}(\nu_{m(i)}, \zeta_{m(i)}), \quad \text{for } m(i) = \text{standard},$$

where

$$\nu_i \sim \mathcal{N}(1, 0.5),$$

and

$$\zeta_i \sim \mathcal{G}(4, 1).$$

These priors imply the coverage random effects characteristics to be as in Table 3. This (subjective) specification is based on the experiences in the MIMOSA project.

Table 3: Characteristics of the priors for random parameters

	$q(0.25)$	median	$q(0.75)$
standard coverage	0.26	0.50	0.74

Finally, for the constants in the migration models, a normal hierarchical prior is assumed

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

The same priors for the constants in the model for flows amongst EU/EFTA countries and the rest of world are used. 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, as well as precision of the flow-specific random effects U, we assume a gamma prior density $\mathcal{G}(1, 1)$.

6 Results

The IMEM model was developed in the OpenBUGS software dedicated to Bayesian computations. The posterior characteristics were computed with MCMC samples of 1,000 with the 200,000 burn-in sample. The MCMC is a numerical method allowing us to simulate samples, which, in turn are used to reconstruct the marginal posterior density of any parameter of interest.

6.1 Goodness of fit

In order to assess the goodness of fit of the model the posterior predictive p-values for all available data were computed (Gelman et al. 2004). The predictive distributions (replicated data) for all $\hat{z}^k_{ijt}$ were computed as

$$p(\hat{z}|z) = \int p(\hat{z}|\theta)p(\theta|z)d\theta. \tag{14}$$

Technically, a 1000 element sample from a posterior distribution of each of the parameters was saved. Then, the $\log \bar{\mu}^k_{ijt}(n)$, $n = 1, \ldots, 1000$, was computed as

$$\log \bar{\mu}^S_{ijt}(n) = \log y_{ijt}(n) + \psi_i(n) - \log\left(1 + e^{-\kappa_i(n)}\right), \tag{15}$$

$$\log \bar{\mu}^R_{ijt}(n) = \log y_{ijt}(n) + \gamma_j(n) - \log\left(1 + e^{-\kappa_j(n)}\right). \tag{16}$$

In the next step, $\log \mu^k_{ijt}(n)$ were drawn from normal distributions

$$\log \mu^k_{ijt}(n) \sim \mathcal{N}(\log \bar{\mu}^k_{ijt}(n), \varepsilon^k_{ijt}(n)).$$

Then, a 1000 replication of the predicted data were drawn

$$\hat{z}^k_{ijt}(n) \sim \mathrm{Po}(\mu^k_{ijt}(n)).$$

The posterior predictive p-values were computed as a ratio of the replicated data exceeding the observed data for each z^k_{ijt}; that is

$$p^k_{ijt} = \frac{\#(\hat{z}^k_{ijt}(n) > z^k_{ijt}(n))}{1000}.$$

If we compute the posterior predictive interquartile range for each of the data points, we should expect that in the well-fitted model half of the data points will lie outside the ranges and half will be inside. Similarly, if we compute, say, 90% ranges around the median of predictive posteriors, we expect the 10% of the observed data to lie outside and 90% to lie within. Thus, we expect all p-values to be uniformly distributed. If the distribution is concave (downwards) then less data points lie in the extremes of their predictive posteriors, which suggests that the model is overfitted, that is, we are 'pessimistic' about the uncertainty. The reverse situation, that is the convex (downwards) distribution of the p-values suggests relatively worse fit to the data. It may mean that the uncertainty is too small or that the patterns (spatial or time) are not well reflected by the model. Note, that inference about the magnitude of the goodness of fit should not be based upon the p-values.

Figure 1 shows the cumulative distribution of the p-values for observed emigration and immigration data and Figure 2 shows the histogram of all p-values. We observe that the shape of distributions is concave, which means that the model overfits the observed data. This situation is not desired but we believe that it is better to provide too wide spans of uncertainty than too narrow. However, the distribution is symmetric, which means that the predictions based on the model are equally likely to lie above or below the observed values. In Figure 3 distributions of p-values for three types of the uncertainty are presented. We observe that for all three types there is a predominance of the middle values but it is clearest for the Scandinavian countries with the best data collection systems. For them, the extreme p-values are comparatively rare. This means that the observed data too often lie within the posterior distribution of the true flows, hence, the uncertainty of these flows is too large.

Figure 1: Posterior densities of the Bayesian p-values

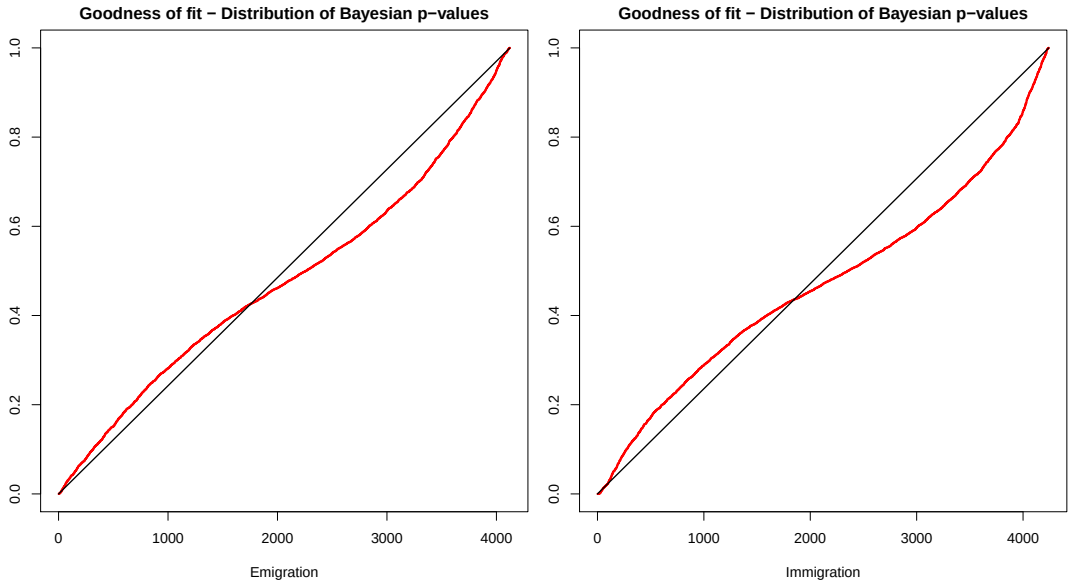
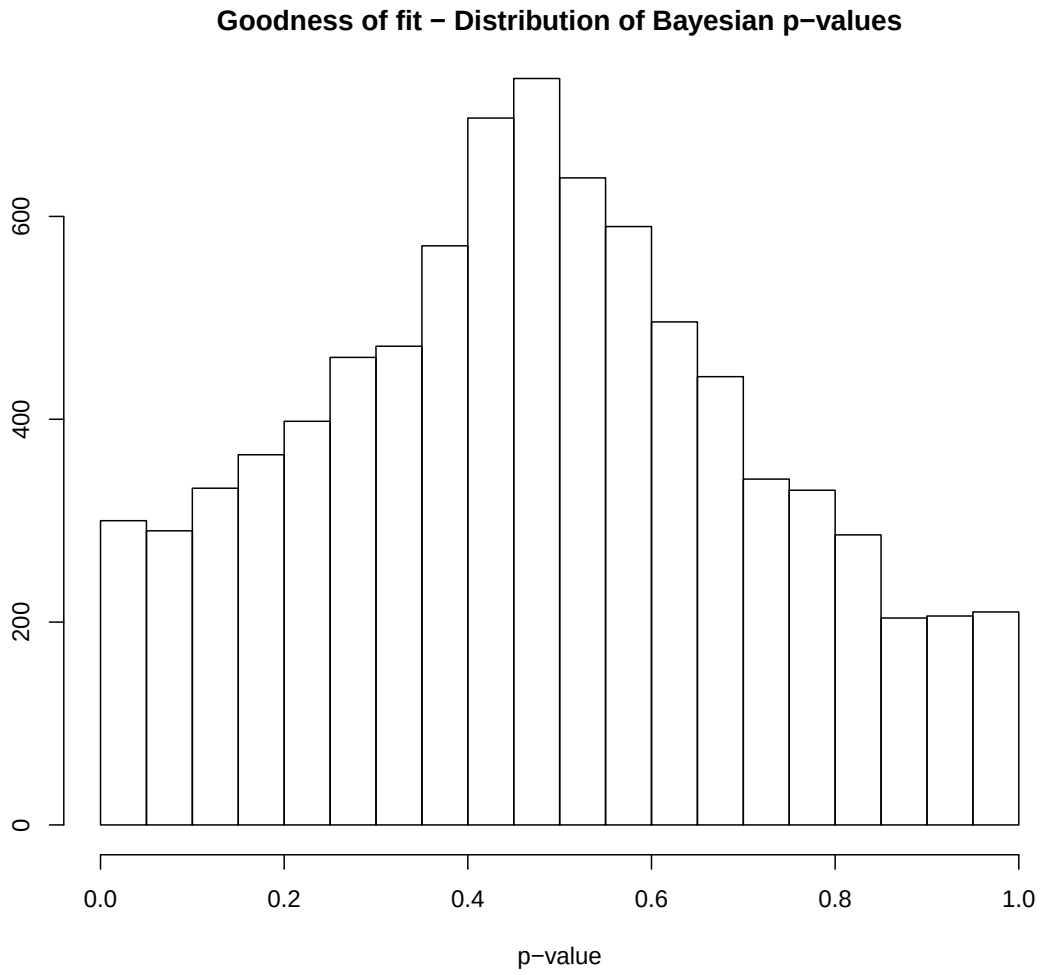


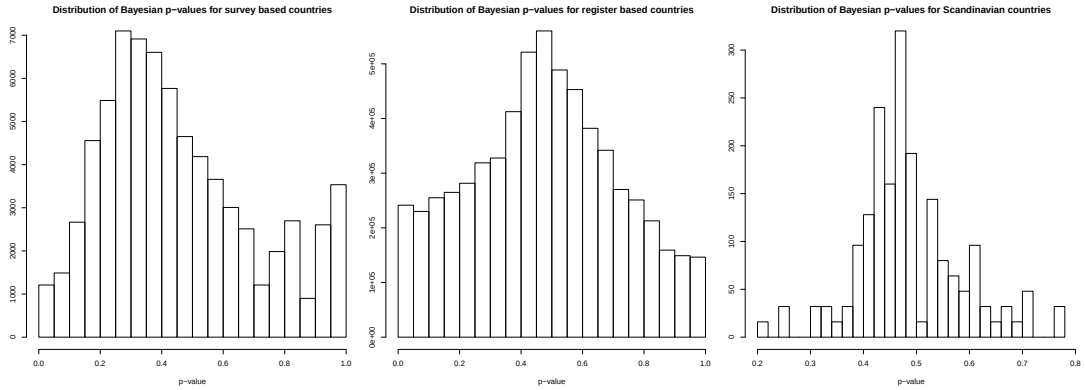
Figure 2: Posterior densities of the Bayesian p-values



Another approach for assessing the goodness of fit of the model is to compare the absolute errors of the

To assess how the model predicts data on emigration and immigration, we com-

Figure 3: Posterior densities of the Bayesian p-values



pared the means of the posterior distributions of the estimators of the observed data ($\hat{z}^k_{ijt}(n)$ computed as above) with the available data. Table 4 presents the mean absolute errors and median absolute errors for all 31 countries and seven years for the intra European flows. The sums of flows are given for comparison. We observe that the model predicts the data reasonably well.

Table 4: Prediction of the data by the model

	Intra EU/EFTA		Rest of world	
data	Mean AE	Median AE	Mean AE	Median AE
Emigration	4,306	14	122,785	5,512
Immigration	309	5	51,419	7,596

Source: own computations

6.2 Model results

The posterior means of the duration of stay factors, which are defined as $\exp(-\delta_m)$, are presented in Table 5. We can interpret them as a multiplicative adjustment 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 73% of the observed data. For six months, this factor is 95%. For permanent duration the true flows are on average thrice larger than the observed data. The posterior densities of the duration of stay factors are presented in Figure 4.

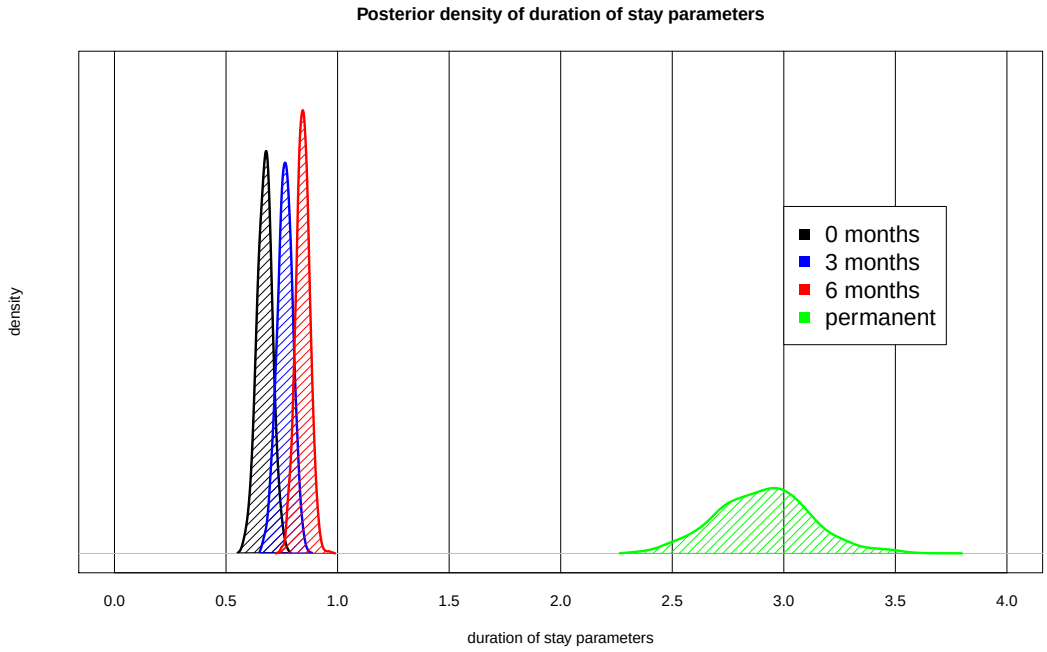
Table 5: Posterior characteristics of the duration criteria factors

duration	no time limit	3 months	6 months	**12 months**	permanent
mean	67%	76%	84%	**100**%	291%
std.dev	3.3%	3.4%	3.0%	NA	21.2%

Source: own computations

Within the EU and EFTA, the undercount of flows is estimated to be 62% on average (standard deviation 5.9%) for emigrants and 77% on average (with standard deviation 7.1%) for immigrants. That is, the observed emigration data constitute 62% and 77%, respectively, of the true flows. The estimated undercount of emigration to the rest of the world is estimated to be 38% of the true flows (with standard deviation 11.1%). The corresponding figure for immigration from the rest of world is 51% (with standard deviation 11.5%). As the identification of undercount parameters

Figure 4: Posterior densities of the duration criteria parameters



are not possible from the data alone, these expert-based priors were particulary informative for our model.

Accuracy of the data collecting systems is measured by precision of the error terms in the measurement equations. The posterior means are presented in Table 6. As expected, we observe that measurement of immigration is always more accurate than emigration. Secondly, the most accurate are the Scandinavian countries, then the other countries with register based systems and the least accurate are the countries using surveys to collect data.

Table 6: Posterior characteristics of the precision parameters

Accuracy type	mean	std.dev	median
Emigration			
Scandinavia	0.16	0.019	0.16
Registers	2.04	0.082	2.04
Surveys	594.6	553.7	582.3
Immigration			
Scandinavia	0.42	0.077	0.40
Registers	5.12	0.294	5.10
Surveys	232.5	251.0	131.9

Source: own computations

The ultimate aim of this research is to produce posterior distributions of all the true flows amongst the 31 countries from 2002 to 2008. In Table 7 and Figure 5, we present posterior characteristics and densities of the 2006 flows from Denmark to The Netherlands and from France to Hungary. For the Denmark to The Netherlands flow, both countries provided data, resulting in a prior that is comparatively tight (relative deviation is 24%), with mean around 630 people. For the flow from France to Hungary, on the other hand, neither country provided data. Here, the posterior density is based primarily on the migration model. This flow is characterised by a

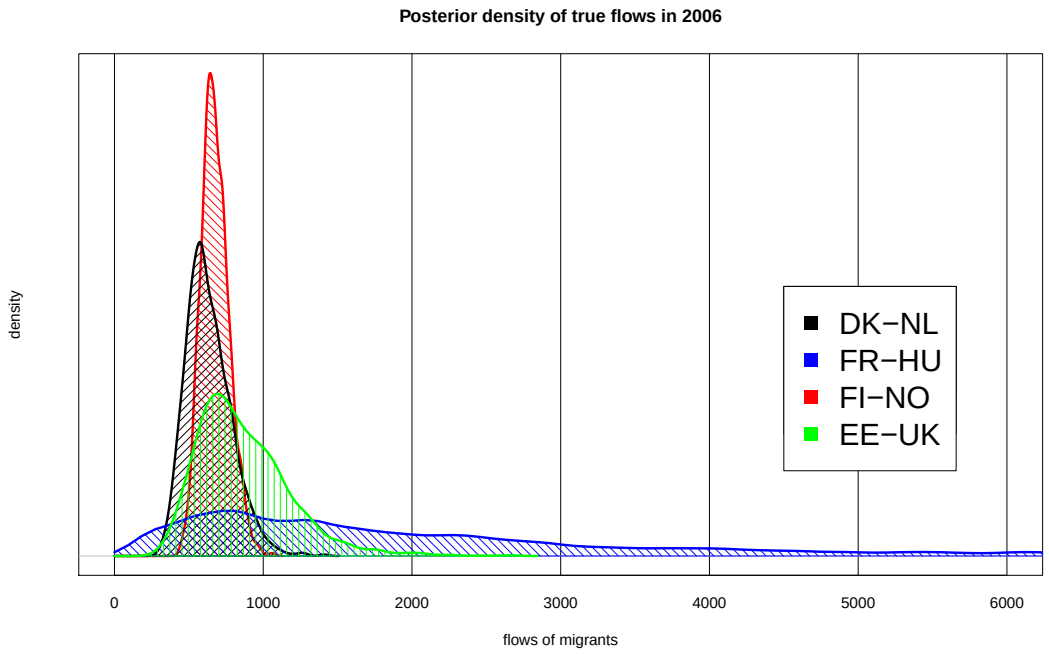
relatively large amount of uncertanty and a heavy right tail, the relative error is enormous – 116%. Flow from Finland to Norway is estimated to be 676 people in 2006. It is characterised by relatively high precision (14%), which results from the fact that these countries exchange their data on migrations. The last presented flow, from Estonia to the United Kingdom, is a bit more uncertain, a mean of 867 people and standard deviation of 311 (relative deviation is 36%). Despite having both pieces of information about this flow, the UK data is assumed to be inaccurate due to the sampling error.

Table 7: Posterior characteristics of the migration true flows

flow	mean	std.dev	rel.std.dev	median
DK-NL	634	152	24%	612
FR-HU	2,554	2,971	116%	1,625
FI-NO	676	93	14%	668
EE-UK	867	311	36%	812

Source: own computations

Figure 5: Posterior densities of the migration true flows



As another illustration, consider the 2006 flows from Poland to Germany and from Finland to Sweden presented in Table 8 and Figures 6 and 7, respectively. The posterior true flow from Poland to Germany (Figure 6) is 167,500 people with standard deviation about 40,200. Here, the reported data, also shown in the figure, differ considerably from our estimated true flows. In fact, they are both lower than the mean or median of the distribution. The reason for this has to do with Poland and Germany's duration of stay criteria used to identify migrants. Poland uses a permanent duration, which results in a relatively small number of migrants recorded (around 15,000). This number is additionally lowered in our model due to the undercount of emigration. In the German data collection system no time limit is applied for incoming flows. This results in a relatively large number of immigrants (around

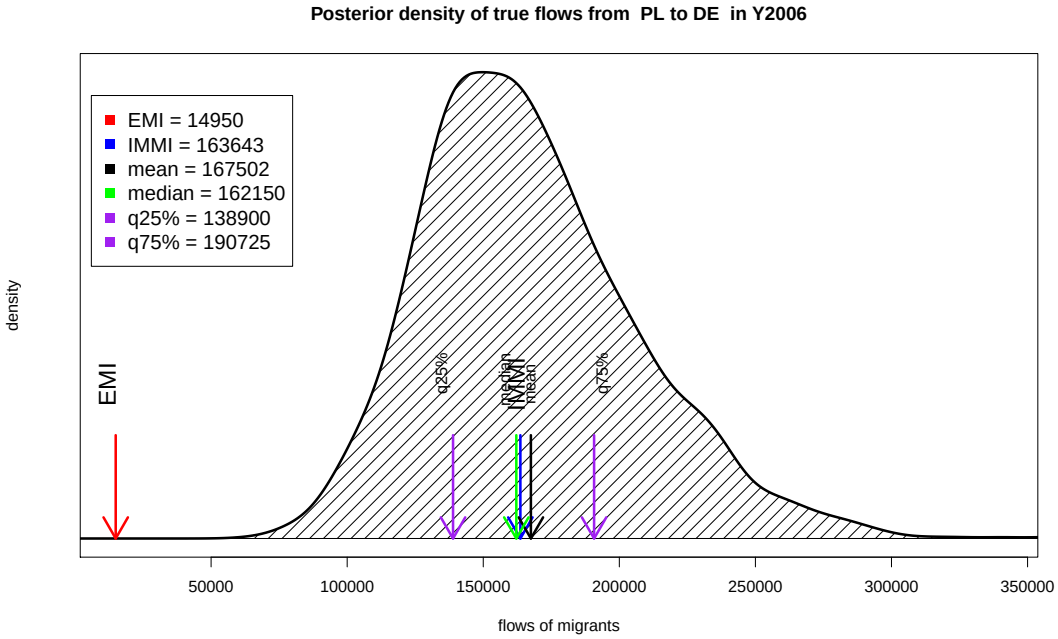
164,000), but the undercount of immigration, indicated by the experts, seems to balance the effect of the shorter duration criterion.

Table 8: Posterior characteristics of the migration true flows

flow	mean	std.dev	median
PL-DE	167,500	40,280	162,200
FI-SE	4,086	527	4,086

Source: own computations

Figure 6: Posterior densities of the migration true flows

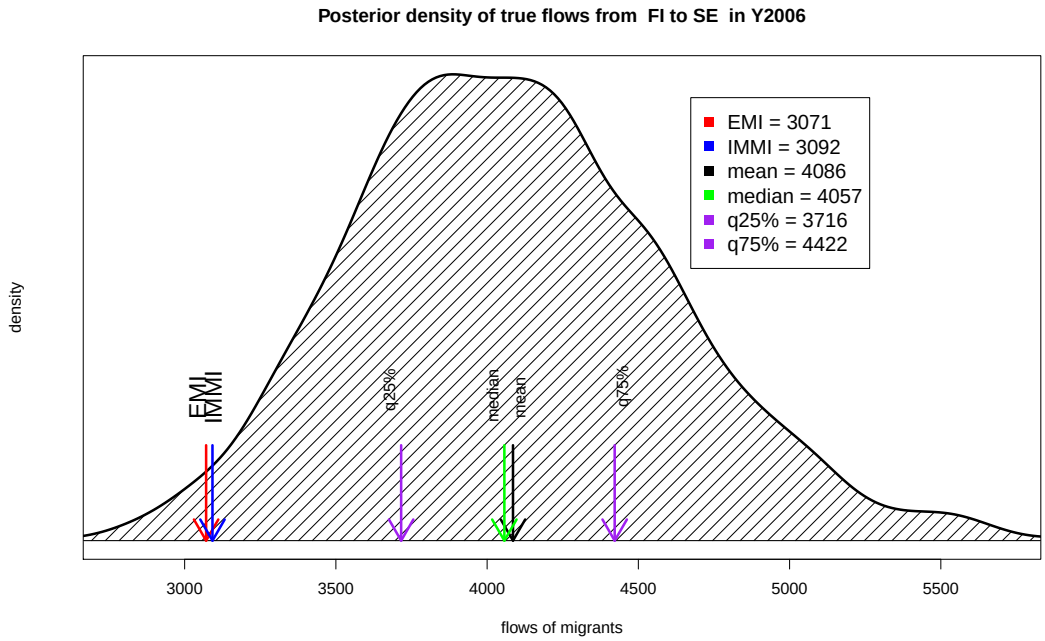


In Figure 7, a posterior density of the 2006 migration flow from Finland to Sweden is presented. The mean is around 4,100 migrants with standard deviation of around 530. We also observe that the data reported by both sending and receiving countries are very close to each other (around 3,000). Both reported flows lie in a tail of the posterior density and they are significantly lower than the mean or median of the posterior true flow. The reason for this is due to our inclusion of expert information on the undercount of immigration and emigration and a very high precision of the estimate (note, the Scandinavian countries exchange information about the migration statistics). In the MIMOSA project, Sweden's immigration data represented the benchmark and was assumed to be measured without error or undercount. In the IMEM model, however, the subjective expert assessment of the immigration undercount by means of a prior density on λ_2 is incorporated. This leads to higher average flows than reported by the receiving countries.

7 Conclusions

In this paper, we have presented the IMEM model, which brings together empirical data, covariate information and expert judgements in a Bayesian modelling framework to estimate migration flows amongst 31 countries in Europe from 2002

Figure 7: Posterior densities of the migration true flows



to 2008. The covariate information was used to estimate missing flows. The expert judgements are used to inform the measurement model and to overcome the limitations in the existing data. While there is still some work to be done to improve the estimates and measures of uncertainty, overall we believe the model is producing reasonable and informative results. The next steps for this research are to include more detail concerning the accuracy of reported migration flows and to extend the model to include age and sex.

To conclude, we hope this work provides an important foundation for both modelling and understanding international migration, particularly in situations where the data are inadequate or missing. We have shown how data obtained from multiple sources with different measurements and collection systems can be combined together to provide a more complete and consistent picture of international migration.

Acknowledgements

This research is part of the Integrated Modelling of European Migration (IMEM) project funded by the New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE). The authors would like to thank the migration data experts for providing their judgements, as well as the other IMEM team members, Guy J. Abel, Solveig Christiansen, Nico Keilman and Rob van der Erf, for their comments and suggestions on this work.

References

[1] Abel GJ (2010) Estimation of international migration flow tables in Europe. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 173(4):797-825.

[2] Bijak J (2011) *Forecasting international migration in Europe: A Bayesian view.* Dordrecht: Springer.

[3] Bijak J and Wiśniowski A (2010) Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 173(4):775-796.

[4] Bilsborrow RE, Graeme H, Amarjit SO, and Zlotnik H (1997) *International migration statistics: Guidelines for improving data collection systems.* Geneva: International Labour Office.

[5] 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.

[6] Chunn JL, Raftery AE and Gerland P (2010) Bayesian probabilistic projections of life expectancy for all countries. Working Paper no. 105, Center for Statistics and the Social Sciences, University of Washington.

[7] Cohen JE, Roig M, Reuman DC and GoGwilt C (2008) International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences* 105(40):15269-15274.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Gabrielli D, Strozza S and Todisco E (2009) Country Report Italy. National Data Collection Systems and Practices. PROMINSTAT report.

[14] Girosi F and King G (2008) *Demographic forecasting.* Princeton: Princeton University Press.

[15] González-Ferrer A (2009) Country Report Spain. National Data Collection Systems and Practices. PROMINSTAT report.

[16] 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.

[17] 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.

[18] 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.

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

[20] Kelly JJ (1987) Improving the comparability of international migration statistics: Contributions by the Conference of European Statisticians from 1971 to date. *International Migration Review* 21:1017-1037.

[21] Kim K and 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.

[22] Kupiszewska D and Nowok B (2008) Comparability of statistics on international migration flows in the European Union. In *International migration in Europe: Data, models and estimates*, Raymer J and Willekens F, eds., pp. 41-71. Chichester: Wiley.

[23] 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.

[24] Lynch SM (2007) *Introduction to applied Bayesian statistics and estimation for social scientists.* New York: Springer.

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

[26] Mitchell J, Pain N and Riley R (2011) The drivers of international migration to the UK: A panel-based Bayesian model averaging approach. Paper presented at the Migration: Economic Change, Social Challenge Conference, University College London.

[27] Nowok B (2010) *Harmonization by simulation: A contribution to comparable international migration statistics in Europe.* Amsterdam: Rozenberg Publishers.

[28] 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.

[29] Nowok B and Willekens F (2011) A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place* 17(5):521-533.

[30] 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.

[31] Poulain M (1993) Confrontation des statistiques de migration intra-europd'ż˝ennes: vers une matrice complète? *European Journal of Population* 9(4):353-381.

[32] 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.

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

[34] 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.

[35] 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.

[36] Raymer J and Bijak J (2009) Report of the technical consultancy in the United Kingdom. MIMOSA Deliverable 10.1A, Modelling of Statistical Data on Migration and Migrant Populations, Eurostat Project 2006/S 100-10667/EN LOT 2, Eurostat, Luxembourg.

[37] Raymer J, de Beer J, van der Erf R (2011) Putting the pieces of the puzzle together: Age and sex-specific estimates of migration amongst countries in the EU/EFTA, 2002-2007. *European Journal of Population* 27(2):185-215.

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

[39] 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.

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

[41] van der Erf R (2010) Initial assessment of th quality of international migration data. IMEM Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

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

[43] Warren R and Peck JM (1980) Foreign-born emigration from the United States: 1960 to 1970. *Demography* 17(1):71-84.

[44] 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.

[45] Wiśniowski A, Keilman N, Bijak J, Forster JJ, Smith PWF, Christiansen S and Raymer J (2011) Augmenting migration statistics with expert knowledge. Paper prepared for IMEM Workshop, Chilworth, 25-27 May.

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

[47] Zaba B (1987) The indirect estimation of migration: A critical review. *International Migration Review* 21(4):1395-1445.

[48] Zlotnik H (1987) The concept of international migration as reflected in data collection systems. *International Migration Review* 21(4):925-946.