

# Integrated Modeling of European Migration

James Raymer\*, Arkadiusz Wiśniowski,

Jonathan J. Forster, Peter W.F. Smith and Jakub Bijak

Southampton Statistical Sciences Research Institute

University of Southampton

March 20, 2013

## Abstract

International migration data in Europe are collected by individual countries with separate collection systems and designs. As a result, reported data are inconsistent in availability, definition and quality. In this paper, we propose a Bayesian model to overcome the limitations of the various data sources. The focus is on estimating recent international migration flows amongst 31 countries in the European Union and European Free Trade Association from 2002 to 2008, using data collated by Eurostat. We also incorporate covariate information and 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.

Key words: international migration statistics, migration models, uncertainty, Europe, Bayesian modeling

---

\*Contact email at raymer@soton.ac.uk

# 1 INTRODUCTION

In order to fully understand the causes and consequences of international population movements, researchers and policy makers need to overcome the limitations of the various data sources countries use to produce statistics, including inconsistencies in the availability, definitions and quality (Kelly 1987; Zlotnik 1987; Willekens 1994; Bilborough et al. 1997; Poulain et al. 2006; Kupiszewska and Nowok 2008). In this paper, we propose a Bayesian model for harmonizing and correcting the inadequacies in the available data and for estimating the completely missing flows, where harmonizing refers to the process of reconciling the differences between various measurements of migration data. 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 collated by Eurostat that are based on reports by national statistical offices. The methodology is integrated and capable of providing a synthetic data base of estimates with measures of uncertainty for international migration flows and other model parameters.

The advantages of 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 increasingly the major factor contributing to population change (Goldin et al. 2010), especially for countries in Europe already experiencing declines in their working age populations (Castles and Miller 2009, pp. 223-224). Furthermore, our understanding of how or why populations change necessitates reliable information about migrations. Finally, countries in Europe are now required to provide harmonized migration flow statistics to Eurostat as part of a new Regulation passed by the European Parliament in 2007 (No. 862/2007). Recognizing 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.’ Our proposed framework helps countries achieve this aim and provides measures of accuracy for the estimated parameters and flows.

The paper is structured as follows. In Section 2, we first provide some background

on the problems and inconsistencies with available data on migration flows. In Section 3, we present our methodology for estimating international migration in Europe that integrates data, knowledge on differences in measurement, expert-based judgement and covariate information. Our results and assessment of the model are presented in Section 4. Finally, the paper ends with some conclusions in Section 5.

## 2 BACKGROUND

The reasons for international migration are many. People move for employment, family reunion or amenity reasons. Reported statistics on population flows, on the other hand, are relatively confusing or nonexistent. There are two main reasons. First, no consensus exists on what exactly constitutes a ‘migration’. Therefore, comparative analyses suffer from differing national views concerning the definition of 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 by counting changes in residence. The challenge is compounded because countries use different methods for data collection. Migration statistics may come from a variety of administrative registers, censuses or surveys.

The timing (duration) criterion used to identify international migrants varies considerably between countries. For example, in the German register there is no time criterion, i.e., everyone who enters the country not for the purposes of tourism or business is obliged to register and should be counted as an immigrant. On the other hand, in Poland, immigrants are those who become registered for permanent stay in the country.

International migration statistics also suffer from reliability problems, mainly due to under-registration of migrants and imperfect data coverage (Nowok et al. 2006). Under-registration is often caused by migrants not notifying the authorities in charge of the population register of their movement. This is particularly an issue for measuring emigrants, where the persons may have very little incentive to deregister. Surveys, such as the United Kingdom’s International Passenger Survey, often do not have large enough sample sizes to adequately capture the details needed for analyzing migration (De Beer et al., 2010; Raymer et al. 2011a). This is because flows of international migrants only represent a small fraction of any population, and because migrants might

be more difficult to capture than the rest of the population. Finally, data on flows for certain countries may be missing for particular years or even entirely.

Because of all the problems associated with inconsistency and missing data, there has only been a limited amount of work carried out in the area of estimating or forecasting international migration flow tables. Most of this work has been focused on indirect methods for particular countries (e.g., Warren and Peck 1980; Jasso and Rosenzweig 1982; Hill 1985; Zaba 1987; Van der Gaag and Van Wissen 2002; Bijak 2010; Bijak and Wiśniowski 2010). There are, however, several recent papers on harmonizing and estimating migration flow tables from which we can draw experiences. Abel (2010) and De Beer et al. (2010) provide extensions of Poulain's (1993) constrained optimization procedure to minimize the differences between two origin-destination migration flow tables representing sending and receiving country reported statistics. Van der Erf and Van der Gaag (2007) and DeWaard et al. (2012) developed iterative hierarchical procedures to allow countries providing better data to have more weight in the estimation. Nowok (2010) proposed a simulation-based approach (see also Nowok and Willekens 2011). Abel (2012) developed a method for estimating flows based on birthplace-specific migrant stock data obtained from decennial censuses. Finally, Raymer (2007, 2008), Brierely et al. (2008), Cohen et al. (2008), Abel (2010), Kim and Cohen (2010) and Raymer et al. (2011b) developed methods for estimating missing flows. Our approach to harmonizing and estimating migration differs from previous attempts by the emphasis on modeling the measurement aspects of the reported statistics and by providing measures of uncertainty for all flow estimates and parameters in the model. Furthermore, we have come to the conclusion that a Bayesian approach offers the best opportunity for integrating all the different types of data, covariate information and a priori knowledge.

There are two important advantages of adopting a Bayesian approach in the context of estimating international migration flows. First, the methodology offers a coherent probabilistic mechanism for describing various sources of uncertainty contained in the various levels of modeling. These include the migration processes, models, model parameters and prior information. Second, as noted by Willekens (1994), the methodology provides a formal mechanism for the inclusion of expert judgment to supplement the deficient migration data.

### 3 METHODOLOGY

The conceptual framework of the model we develop for estimating international migration flows is presented in Figure 1. The interest is in estimating a set of unobserved true flows of migration based on four pieces of information: flows reported by the sending country, flows reported by the receiving country, covariate information and expert judgments. The reported data are harmonized via two measurement models: one for the sending country data and one for the receiving country data. These models distort the true flows by taking into account duration definitions used in various countries, relative accuracy of the data collection mechanisms, the overall undercount of migration and coverage of migrants. A migration model based on theory is used to augment the measurement model and to estimate the missing flow data. In the following sections, we describe the main design aspects of our methodology: (i) specification of the data model, (ii) the development of the measurement error model, (iii) elicitation of expert-based prior distributions for the measurement model, and (iv) the migration model, which permits estimation in the presence of missing data.

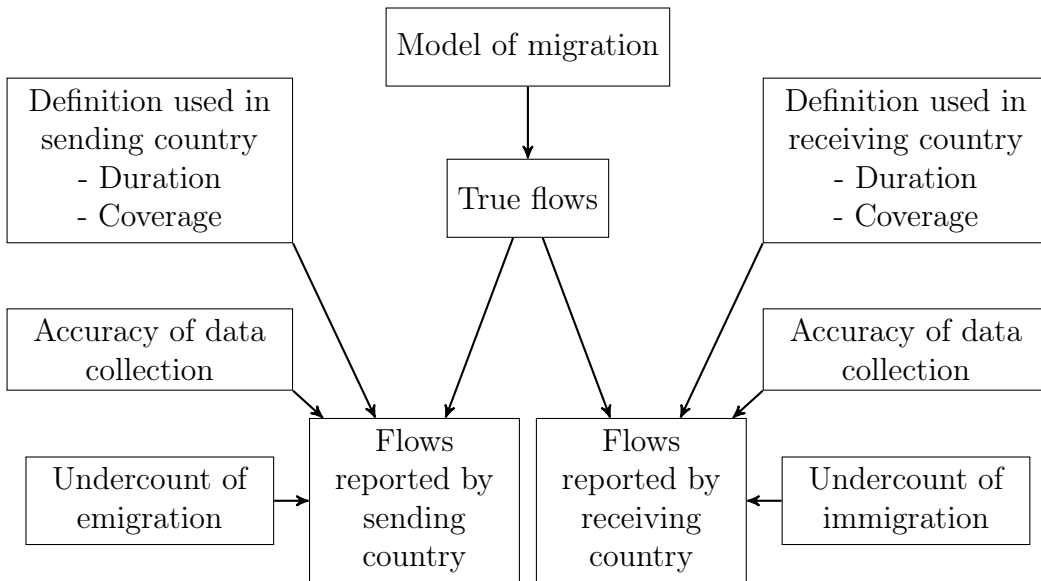


Figure 1: Conceptual framework for modeling migration flows

#### 3.1 Data model

The migration flow data used in the project come primarily from the Eurostat data base, which relies on the annual Joint Questionnaire on Migration Statistics collected from all national statistical agencies in the European Union. This questionnaire is

coordinated by Eurostat, and is sent out on behalf of the Council of Europe, the United Nations Statistical Division, the United Nations Economic Commission for Europe and the International Labour Organization. The migration data from Eurostat represent reported flows amongst the 31 countries in the EU and EFTA, and to and from the rest of world, from 2002 to 2008. In Table 1, these countries are listed along with their population sizes in 2008 and categories of accuracy (A), duration criteria for migration received (D-R) and sent (D-S), and undercount (U). The categories are described in more detail in Section 3.2.

In Figure 2, we present a more detailed specification of our model. The international migration flow data of interest can be expressed in a two-way contingency table or matrix. 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  $k \in \{S, R\}$ . These flows can be represented by matrices  $Z_t^S$  and  $Z_t^R$ :

$$Z_t^S = \begin{pmatrix} 0 & z_{12t}^S & z_{13t}^S & \cdots & z_{1nt}^S \\ z_{21t}^S & 0 & z_{23t}^S & \cdots & z_{2nt}^S \\ z_{31t}^S & z_{32t}^S & 0 & \cdots & z_{3nt}^S \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1t}^S & z_{n2t}^S & z_{n3t}^S & \cdots & 0 \end{pmatrix}, \quad Z_t^R = \begin{pmatrix} 0 & z_{12t}^R & z_{13t}^R & \cdots & z_{1nt}^R \\ z_{21t}^R & 0 & z_{23t}^R & \cdots & z_{2nt}^R \\ z_{31t}^R & z_{32t}^R & 0 & \cdots & z_{3nt}^R \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1t}^R & z_{n2t}^R & z_{n3t}^R & \cdots & 0 \end{pmatrix}.$$

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

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

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

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

Table 1: European Union (EU) and European Free Trade Association (EFTA) Countries: Population sizes (in thousands) and measurement aspects of migration data

Code	Name	Population	A	D-R	D-S	U
AT	Austria	8337	2	2	2	1
BE	Belgium	10667	—	—	—	—
BG	Bulgaria	7623	3	4	4	2
CH	Switzerland	7648	—	—	—	—
CY	Cyprus	793	3	0	0	1
CZ	Czech Republic	10424	3	2	4	2
DE	Germany	82110	2	1	1	1
DK	Denmark	5494	1	3	3	1
EE	Estonia	1341	3	4	4	2
ES	Spain	45556	2, 3	1	1	1,2
FI	Finland	5313	1	0	0	1
FR	France	64167	—	—	—	—
GR	Greece	11237	—	—	—	—
HU	Hungary	10038	—	—	—	—
IE	Ireland	4426	3	0	0	1
IS	Iceland	317	1	3	3	1
IT	Italy	59832	3	1	0	1
LI	Liechtenstein	35	—	—	—	—
LT	Lithuania	3358	3	0	3	2
LU	Luxembourg	489	3	1	1	1
LV	Latvia	2266	3	0	3	2
MT	Malta	412	—	—	—	—
NL	Netherlands	16446	2	3	0	1
NO	Norway	4768	1	3	3	1
PL	Poland	38126	3	4	4	2
PT	Portugal	10622	—	—	—	—
RO	Romania	21514	3	4	4	2
SE	Sweden	9220	1	0	0	1
SI	Slovenia	2021	3	4	2	2
SK	Slovakia	5407	3	4	4	2
UK	United Kingdom	61179	3	0	0	1

Notes: (i) Accuracy (A) refers to migration data system: 1 = Nordic register, 2 = Other good register, 3 = less reliable register or survey, — no country of origin / destination data available; (ii) Durations (D) are specified for receiving (R) and sending (S) countries: 1 = no time limit, 2 = three months, 3 = six months, 0 = twelve months, 4 = permanent; (iii) Undercount (U): 1 = low, 2 = high (see Section 3.2); (iv) Spain has two entries for the A and U columns because the measurement of immigration is considered to be much better than emigration.

### 3.2 Measurement error model

In our model,  $y_{ijt}$  is a true flow of migration from country  $i$  to country  $j$  in year  $t$  (see Figure 2). It includes migration flows to and from the rest of world. In terms of measurement, true flows are consistent with the United Nations (1998, p. 18) recommendation

for long-term international migration, i.e., 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’.

To convert the reported data to comply with the UN definition, we use the following two measurement error equations:

$$\log \mu_{ijt}^S = \log y_{ijt} + \delta_{m(i)} - \log \lambda_{f(i)} - \log (1 + e^{-\kappa_i}) + \varepsilon_{ijt}^S, \quad (1)$$

$$\log \mu_{ijt}^R = \log y_{ijt} + \delta_{m(j)} - \log \lambda_{g(j)} - \log (1 + e^{-\kappa_j}) + \varepsilon_{ijt}^R, \quad (2)$$

where the differences in the duration of stay criterion are captured by  $\delta_{m(i)}$  and the effects of the undercount are captured by  $\lambda_{f(i)}$  and  $\lambda_{g(j)}$ . We assume  $\varepsilon_{ijt}^S \sim \mathcal{N}(0, \tau_{c(i)}^S)$  and  $\varepsilon_{ijt}^R \sim \mathcal{N}(0, \tau_{c(j)}^R)$ . Here, and throughout this paper, we use  $\mathcal{N}(\mu, \tau)$  to denote a normal distribution with mean  $\mu$  and *precision* (inverse variance)  $\tau$ .

The  $\delta_{m(i)}$  parameter measures the effect of a particular minimal duration of stay definition used by country  $i$ , with the following categories included:

$$\delta_{m(i)} = \begin{cases} \delta_1 & \text{if criterion is ‘no time limit’} \\ \delta_2 & \text{if duration is 3 months} \\ \delta_3 & \text{if duration is 6 months} \\ 0 & \text{if duration is 12 months} \\ \delta_4 & \text{if duration is permanent} \end{cases} . \quad (3)$$

The parameters are 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.



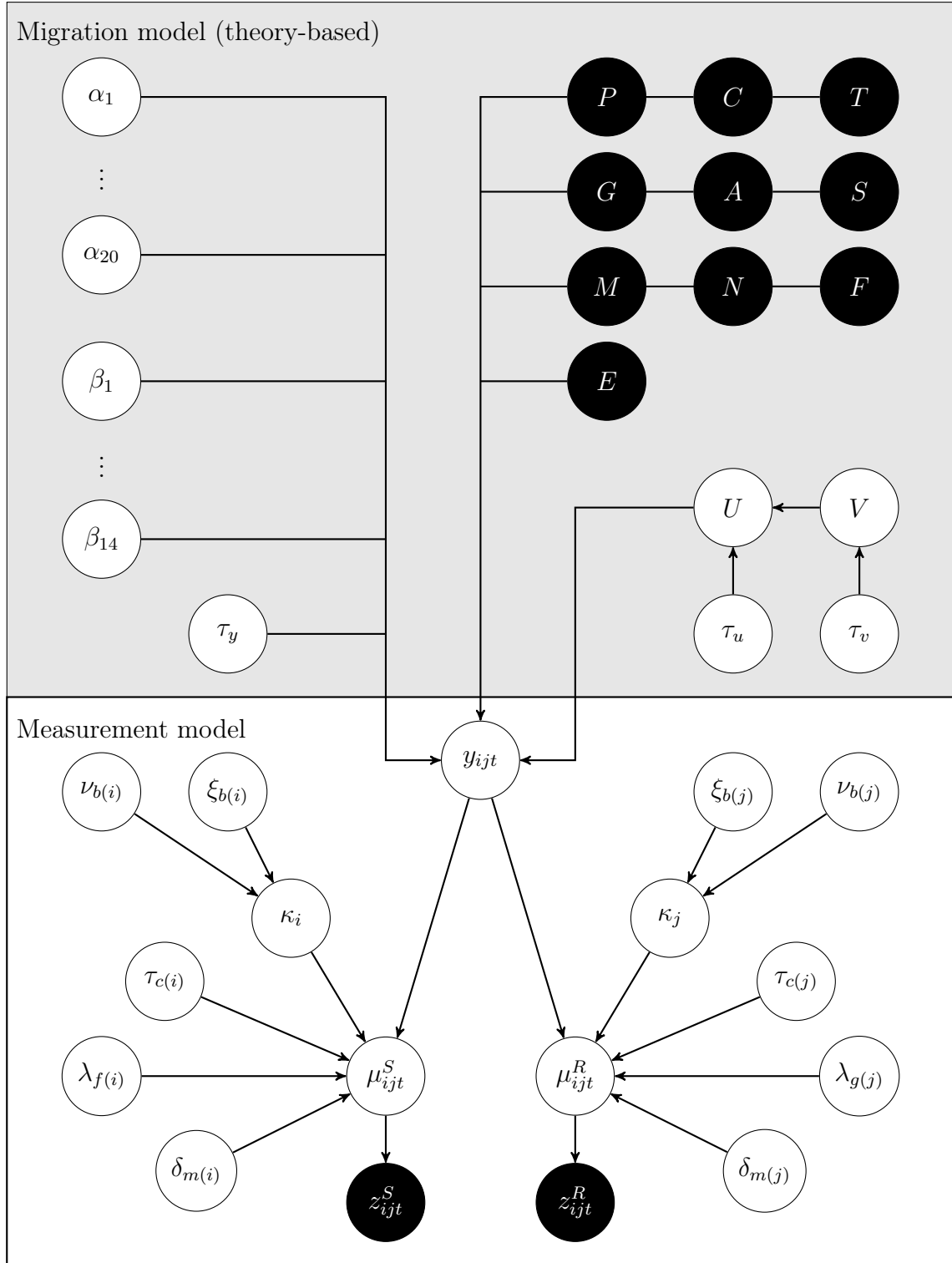


Figure 2: Graphical representation of the integrated model for European migration

The parameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  take values in  $(0, 1)$  and determine the effect of the undercount of emigration or immigration. For each flow, the appropriate  $\lambda$  parameter

is assigned according to the value of  $f(i)$  or  $g(j)$ , where

$$f(i) = \begin{cases} 1 & \text{if undercount of emigration is assumed low} \\ 3 & \text{if undercount of emigration is assumed high,} \end{cases} \quad (4)$$

$$g(j) = \begin{cases} 2 & \text{if undercount of immigration is assumed low} \\ 4 & \text{if undercount of immigration is assumed high.} \end{cases} \quad (5)$$

The classifications of undercount in countries are presented in the last column in Table 1. It is a well-acknowledged fact that the official statistics suffer from underreporting (see, e.g., Bilsborrow et al., 1997; Poulain et al, 2006; Kupiszewska and Nowok, 2008). The undercount particularly affects registers which are based on self-declarations. People may not register or deregister for various reasons, and there may be no requirement to do so. This undercount is deemed to be more severe in the case of emigrants, who usually have fewer incentives to deregister from the system than immigrants who, after registration, may gain access to certain benefits, such as health insurance, education, pension schemes or social benefits. We assume that there is a certain level of undercount of both emigration and immigration in all countries. Furthermore, the data collection systems in the countries under study can be divided into two general groups: ‘low’ and ‘high’ undercount. The classification relies on our own expertise, as well as assessments of the data collection systems in Europe obtained from various studies (see Poulain et al., 2006; Kupiszewska and Wiśniowski 2009; Van der Erf 2009).

The  $\kappa_i$  parameter is a normally distributed country-specific random effect,  $\kappa_i \sim \mathcal{N}(\nu_i, \zeta_i)$ , where  $\nu_i = \nu_{b(i)}$  is a group-specific mean,  $\zeta_i = \zeta_{b(i)}$  is a group-specific precision and  $b(i)$  denotes a type of coverage assumed for country  $i$ . We assume two coverage types, i.e.,  $b(i) \in \{\text{standard, excellent}\}$ . Moreover, we assume that the coverage is same when measuring emigration and immigration apart from the registers in Italy, Romania and Spain, which have large discrepancies between their measurement processes of emigration and immigration. For instance, in Romania, the reported immigration only includes foreigners, while the reported emigration only includes nationals. The logistic transformation of  $\kappa_i$  in (1) and (2) ensures that the random effect is within the range  $(0, 1)$  on the linear scale. This parameter captures the country-specific deficiencies of the data collection system in measuring migrants which are not reflected by the overall

undercount  $\lambda$ , and can be interpreted as the difference in coverage with respect to the United Nations definition of migration. For the five Nordic countries and The Netherlands, this coverage is constrained to be excellent (i.e., it is set to 1 on the linear scale) which ensures identifiability of the random effects.

Finally, the variances of the error terms depend on whether the data are captured by sending or receiving countries, respectively, and the type of collection system,  $c(i)$ . 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, undercount and coverage. As shown in the fourth column of Table 1, we distinguish three types of data collection systems for migration flows: (i) registers in the the Nordic countries which exchange information on migration flows, (ii) other good register-based systems and (iii) less reliable register-based or survey systems. The countries not reporting any migration data by country of origin or destination are Belgium, France, Greece, Hungary, Lichtenstein, Malta, Portugal and Switzerland.

For the migration to and from the rest of world (country 0, denoted also as RW) there is only one equation per outflow and inflow:

$$\log \mu_{i0t}^S = \log y_{i0t} + \delta_{m(i)} - \log \lambda_{f(i)} + \varepsilon_{i0t}^S, \quad (6)$$

$$\log \mu_{0jt}^R = \log y_{0jt} + \delta_{m(j)} - \log \lambda_{g(j)} + \varepsilon_{0jt}^R. \quad (7)$$

All other parameters remain as described above. Note, that in the measurement of the flows to and from the rest of world, perfect coverage is assumed for all countries, i.e., there are no country-specific random effects. This can be justified by the more rigorous registration requirements for migrants originating from or departing to countries outside the EU/EFTA system. To better capture the level of migration, for the flows to and from the rest of world, we have also distributed the category ‘Unknown’ in the reported data proportionally the observed flows.

### 3.3 Expert-based prior distributions

Since it would be impossible to find experts with knowledge about all the 992 flows within our system, we sought information on the overall effects of measurement from 11

experts by means of a Delphi survey. The experts were recruited amongst specialists in European migration data (rather than migration per se) and, in terms of background, represented official statistics, as well as academia. This largely conforms to some of the key principles of the Delphi method (Rowe and Wright 2001), such as the required domain knowledge of the experts, heterogeneity of the panel, and the ideal number of panelists between five and 20.

The Delphi survey followed the example of a similar, migration-related endeavor (Bijak and Wiśniowski 2010), and consisted of a two-round process, with anonymized first-round feedback to the experts provided before the second round. In principle, this provides an opportunity for the experts to converge on views. Although we did not observe much convergence amongst our experts, the second round, nonetheless, proved instrumental in ensuring a shared understanding of the underlying concepts. Note, achieving convergence was never the primary aim of our Delphi exercise. Rather, once common understanding was reached, the remaining differences between the experts constituted yet another source of uncertainty, which was propagated into the model along with the uncertainty about other parameters of the migration and measurement models.

Questions in the Delphi exercise concerned probabilities. For instance, we asked experts to provide a range for a magnitude of duration of stay and to state how certain they were about this range. This allowed us to construct a probability density representing beliefs of each expert. Then, prior distributions for the duration of stay parameters, precision and undercount of the measurement model were created as mixtures of the expert-specific densities. The elicitation of the undercount parameters produced prior distributions with inexplicably high uncertainty. As the undercount of immigration cannot be identified purely from the data, more informative prior distributions were required. Therefore, we decided to use prior distributions elicited from an expert within our team. This expert has detailed knowledge about highly advanced population registers in Europe and the quality of the migration data obtained from them. The whole prior elicitation process, including questions, descriptions of how the answers were transformed into single expert-specific densities and our assessment of the results are discussed in [Placeholder reference].

For the duration of stay parameters,  $\delta_1$ ,  $\delta_2$ ,  $\delta_3$  and  $\delta_4$ , we applied a mixture of

log-normal prior distributions for the auxiliary parameters  $d_m$ , obtained from the experts. For reasons of interpretation, we present  $\exp(-\delta_m)$ . The prior medians for the four parameters (with interquartile ranges in brackets) are:  $\exp(-\delta_1) = 0.51$  (0.39, 0.60),  $\exp(-\delta_2) = 0.61$  (0.50, 0.70),  $\exp(-\delta_3) = 0.81$  (0.73, 0.88) and  $\exp(-\delta_4) = 1.64$  (1.24, 3.55). We can interpret them as multiplicative adjustment factors in the equation *true flow* = *factor*  $\times$  *data*. For example, the median of six months duration parameter,  $\exp(-\delta_3)$ , is equal to 0.81. This value implies that the median of the true flow, measured by using a 12 month duration, would be 81% of the corresponding reported flow measured with a six month duration criterion.

The prior distributions for the precisions of the error terms were also obtained from the experts. More specifically, we collected information on the overall accuracy of all register-based systems, regardless of type (described above). Then we combined the single expert prior distributions into mixtures of gamma densities for the reciprocal variance (i.e., precision) and assumed the same prior distribution for each type of accuracy (see Table 1). Due to the heterogeneity of expert judgments, the resulting prior distributions are rather vague with interquartile ranges of (26, 910) for emigration (median=573) and (171, 1240) for immigration (median=780). A priori, these values imply that experts, not surprisingly, consider emigration to be measured less accurately than immigration. We assume independence in the prior distributions for the precisions of emigration and immigration because it permits the data to provide evidence concerning which type of report (sending or receiving) is more accurate. This information was also elicited separately from the experts.

The prior distributions for the undercount parameters  $\lambda_{f(i)}$  and  $\lambda_{g(j)}$  are beta densities. The medians for the four prior distributions (with interquartile ranges in brackets) are:  $\lambda_1 = 0.73$  (0.69, 0.77),  $\lambda_2 = 0.88$  (0.85, 0.91),  $\lambda_3 = 0.45$  (0.41, 0.50) and  $\lambda_4 = 0.68$  (0.64, 0.72). For example, the median for  $\lambda_1$ , 0.73, implies that for the data collection systems in countries with low undercount, about 73% of the emigration is reported. For emigration from high undercount countries, the median figure is 45%.

The coverage random effects parameters,  $\kappa_i$ , for countries with excellent coverage (Denmark, Finland, Iceland, The Netherlands, Norway and Sweden) are assumed fixed and equal to zero on the logarithmic scale. Hence, the resulting scaling factor for the true flows is equal to one. For the rest of the countries, with standard coverage  $b(i)$ ,

we assume the following:  $\kappa_i \sim \mathcal{N}(\nu_{b(i)}, \zeta_{b(i)})$ , where  $\nu_{b(i)} \sim \mathcal{N}(1, 0.5)$  and  $\zeta_{b(i)} \sim \mathcal{G}(4, 1)$ . These prior distributions have a median coverage random effect of 0.50 with the 25th and 75th percentiles being 0.26 and 0.74, respectively. The same prior distributions are assumed for the emigration- and immigration-specific random effects for Italy, Romania and Spain.

### 3.4 Migration model

The true flows of migration are modeled by using a set of covariates (see Figure 2). Here, we started with Jennissen (2004), Abel (2010) and Raymer et al. (2011b) to gather a set of variables based on migration theories and empirical evidence. Our model for estimating the true flows of migration amongst EU/EFTA countries is specified as

$$\begin{aligned} \log y_{ijt} = & \alpha_1 + \alpha_2 \log P_{it} + \alpha_3 \log P_{jt} + \alpha_4 C_{ij} + \alpha_5 \log T_{ijt} + \alpha_6 \log(G_{it}/G_{jt}) \\ & + \alpha_7 A_{ijt} + \alpha_8 A_{it} + \alpha_9 A_{jt} + \alpha_{10} S_{ij} + \alpha_{11} S_{ji} + \alpha_{12} E_2 + \alpha_{13} E_3 + \alpha_{14} E_4 \\ & + \alpha_{15} E_5 + \alpha_{16} E_6 + \alpha_{17} E_7 + \alpha_{18} N_{ij} + \alpha_{19} M_{ijt} + \alpha_{20} F_{ijt} + U_{ij} + \xi_{ijt}, \end{aligned} \quad (8)$$

where  $\alpha = (\alpha_1, \dots, \alpha_{20})'$  is a vector of parameters. The random term  $\xi_{ijt}$  is assumed to be normally distributed with zero mean and constant precision  $\tau_y$ .

The model above contains the following set of covariates:

1. The mid-year populations in sending and receiving countries, denoted as  $P_{it}$  and  $P_{jt}$ . Source: NewCronos database of Eurostat.
2. Indicator variable for contiguity (or neighboring countries) with 1 if countries  $i$  and  $j$  have a common border and 0 otherwise,  $C_{ij}$ . Source: Mayer and Zignago (2006). Note that contiguity is assumed among all Scandinavian countries.
3. The ratio of the Gross National Income per capita in sending and receiving countries,  $G_{it}$  and  $G_{jt}$ , respectively. Source: World Bank (2010).
4. International trade between origin and destination countries, expressed as imports in current US Dollars,  $T_{ijt}$ . Source: United Nations Commodity Statistics Database.

5. Three indicator variables for 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 the sending country was in the EU/EFTA in year  $t$ . The third,  $A_{jt}$ , is 1 if the receiving country was in the EU/EFTA in year  $t$ .
6. Migrant stocks by country of birth based on population censuses around the year 2000.  $S_{ij}$  denotes the stocks of migrants born in sending country  $i$  and residing in receiving country  $j$ , whereas  $S_{ji}$  denotes the stocks of migrants born in the receiving country and residing in sending country. The former covariate is introduced in order to capture the pull effects (migrant networks), the latter captures the push effects (source of returning migrants). Source: Parsons et al. (2007). Note, Özden et al. (2011) have produced another set of estimates for the year 2000, along with estimates for census years 1960 to 1990. We decided against using these estimates because they are less well known and do not have clear documentation regarding the methodology. They also do not substantially affect our parameter estimates.
7. An indicator variable equal to one for years in which the workers from country  $i$  have been allowed to freely access the labor market in country  $j$ ,  $M_{ijt}$ . Source: European Commission (2006).
8. An indicator variable capturing the effect of opening the labor markets by the United Kingdom and Ireland to the citizens of the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia and Slovenia in 2004,  $F_{ijt}$ . The variable is equal to 1 for years 2004-2008 for flows from these countries to the UK and Ireland and 0 otherwise.
9. An indicator variable for countries sharing the same language family,  $N_{ij}$ . The typology is based on Lewis (2009). When the languages of the sending and receiving countries stem from the same family (e.g., Spanish, Romanian, French and Italian belong to the Italic Romance family)  $N_{ij}$  is equal to 1 and 0 otherwise.
10. Time effect indicator variables for years 2002 to 2007,  $E_t$ ,  $t = 2, \dots, 7$  to capture the different levels of migration over time. The reference year is 2008.

11. In order to smooth the data over time, flow-specific but constant over time random effects are introduced. They are denoted as  $U_{ij}$ , are normally distributed with mean  $V_{ij}$ , where  $V_{ij} = V_{ji}$ , and have a common precision, i.e.,  $U_{ij} \sim \mathcal{N}(V_{ij}, \tau_u)$ .  $V_{ij}$  are normally distributed with mean zero and precision  $\tau_v$ . This structure induces residual correlation between the flows  $y_{ijt}$  and  $y_{jit}$ , i.e., if a flow in one direction is larger than explained by the covariates above, then we expect the flow in the opposite direction to exhibit similar behavior. Similarly, it induces correlation between the same flow at different time points, thus, providing smoothing across time. It also allows borrowing of strength when flow data are missing.

All non-indicator variables were divided by their means and then transformed to a logarithmic scale. The value of one was added to all migrant stocks to remove zero entries.

For modeling flows to the rest of world, we use a model with additional covariates based on Raymer et al. (2011b):

$$\begin{aligned} \log y_{i0t} = & \beta_1 + \beta_2 \log P_{it} + \beta_3 \log G_{it} + \beta_4 H_i + \beta_5 \log S_{0i} + \\ & \beta_6 \log E_{it} + \beta_7 \log L_{it} + U_{i0} + \xi_{i0t}, \end{aligned} \quad (9)$$

and for flows from the rest of world

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

The errors,  $\xi_{i0t}$  and  $\xi_{0jt}$ , are normally distributed with mean zero and precisions  $\tau_{0S}$  and  $\tau_{0R}$ , respectively. The additional covariates are:

1. An indicator variable 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. (2007).
3. Share of the population older than 65 years,  $E_{it}$ . Source: Population Reference Bureau's World Population Data Sheet 2002-2008.



4. Life expectancy at birth of women in years,  $L_{jt}$ . Source: Population Reference Bureau's World Population Data Sheet 2002-2008.
5. The flow-specific and time-constant random effects,  $U_{ij}$ , which are normally distributed with mean zero and precisions  $\tau_{u1}$  for emigration and  $\tau_{u2}$  for immigration. The purpose of these, analogously to the intra-EU model, is to smooth the predicted flows across time.

The prior distributions in the migration model were set to be weakly informative. For the constant in the migration model, a diffuse normal hierarchical prior was assumed with  $\alpha_1 \sim \mathcal{N}(0, \tau_\alpha)$ ,  $\tau_\alpha = 1/a^2$  and  $a \sim \mathcal{U}(1, 10)$ . The same structure was used for the constants  $\beta_1$  and  $\beta_8$  in the rest of world migration flow models. The uniform hyperprior for the standard deviation of  $\alpha_1$ ,  $\beta_1$  and  $\beta_8$  is rather diffuse, while at the same time avoiding MCMC convergence issues which result from allowing excessive prior dispersion for these parameters. For the rest of the parameters in the migration models, that is  $\alpha_i$ ,  $i = 2, \dots, 20$  and  $\beta_i$ ,  $i = 2, \dots, 7, 9, \dots, 14$ , independent weakly informative normal prior distributions,  $\mathcal{N}(0, 0.1)$ , were assumed. For the precisions  $\tau_y$ ,  $\tau_{0S}$  and  $\tau_{0R}$  in the migration models, we assumed independent weakly informative gamma prior densities  $\mathcal{G}(0.1, 0.1)$ . For the precisions of the random effects in the migration model, that is  $\tau_u$ ,  $\tau_v$ ,  $\tau_{u1}$  and  $\tau_{u2}$ , we assumed independent gamma prior densities  $\mathcal{G}(1, 1)$ . All prior densities in the migration model are summarized below:

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

$$\beta_1 \sim \mathcal{N}(0, \tau_{\beta1}), \quad \tau_{\beta1} = 1/b_1^2, \quad b_1 \sim \mathcal{U}(1, 10),$$

$$\beta_8 \sim \mathcal{N}(0, \tau_{\beta2}), \quad \tau_{\beta2} = 1/b_2^2, \quad b_2 \sim \mathcal{U}(1, 10),$$

$$\alpha_i \sim \mathcal{N}(0, 0.1), \quad i = 2, \dots, 20,$$

$$\beta_i \sim \mathcal{N}(0, 0.1), \quad i = 2, \dots, 7, 9, \dots, 14,$$

$$\tau_y \sim \mathcal{G}(0.1, 0.1), \quad \tau_{0S} \sim \mathcal{G}(0.1, 0.1), \quad \tau_{0R} \sim \mathcal{G}(0.1, 0.1),$$

$$\tau_u \sim \mathcal{G}(1, 1), \quad \tau_v \sim \mathcal{G}(1, 1), \quad \tau_{u1} \sim \mathcal{G}(1, 1), \quad \tau_{u2} \sim \mathcal{G}(1, 1).$$

## 4 RESULTS

The Bayesian model for estimating international migration flows was developed in OpenBUGS (Spiegelhalter et al., 2011). The posterior characteristics were computed with MCMC samples of 1,000 with a 50,000 iteration burn-in sample and a thinning of 100 (i.e., each 100<sup>th</sup> iteration was taken).

### 4.1 Model results

Characteristics of the posterior densities are presented in Table 2 for the measurement model and Table 3 for the migration model. The medians and interquartile ranges (in brackets) of the posterior distributions for the duration of stay factors, expressed as  $\exp(-\delta_m)$ , are 0.53 (0.50, 0.55) for no time limit, 0.63 (0.61, 0.64) for three months, 0.73 (0.71, 0.74) for six months and 2.26 (2.12, 2.38) for permanent. Hence, for countries with a ‘no time limit’ of stay criterion, our median true flows constitute 53% of the observed data. For countries applying a permanent duration, the true flows are on average twice as large as the observed data. The posterior densities of the duration of stay factors are presented in Figure 3.

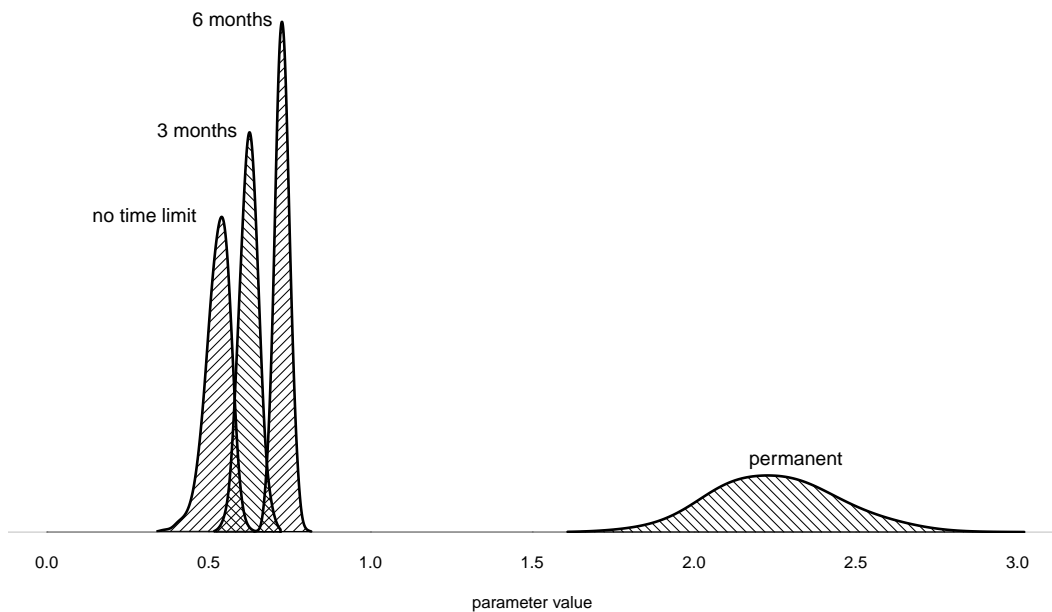


Figure 3: Posterior densities of the duration criteria parameters

Table 2: Posterior characteristics of the measurement model parameters

	parameter	q5%	q25%	median	q75%	q95%
Undercount	$\lambda_1$	0.72	0.75	0.77	0.80	0.82
	$\lambda_2$	0.81	0.84	0.86	0.89	0.92
	$\lambda_3$	0.33	0.36	0.38	0.40	0.43
	$\lambda_4$	0.62	0.67	0.70	0.73	0.78
Duration	$\exp(-\delta_1)$	0.46	0.50	0.53	0.55	0.58
	$\exp(-\delta_2)$	0.58	0.61	0.63	0.64	0.67
	$\exp(-\delta_3)$	0.69	0.71	0.73	0.74	0.76
	$\exp(-\delta_4)$	1.96	2.12	2.26	2.38	2.58
Coverage – random effects	$\text{logit}^{-1}(\kappa_{AT})$	0.77	0.81	0.84	0.87	0.91
	$\text{logit}^{-1}(\kappa_{BE})$	0.09	0.31	0.52	0.73	0.92
	$\text{logit}^{-1}(\kappa_{BG})$	0.05	0.06	0.07	0.07	0.09
	$\text{logit}^{-1}(\kappa_{CH})$	0.11	0.34	0.54	0.74	0.93
	$\text{logit}^{-1}(\kappa_{CY})$	0.63	0.68	0.71	0.76	0.82
	$\text{logit}^{-1}(\kappa_{CZ})$	0.36	0.40	0.43	0.46	0.50
	$\text{logit}^{-1}(\kappa_{DE})$	0.79	0.86	0.91	0.94	0.98
	$\text{logit}^{-1}(\kappa_{EE})$	0.75	0.84	0.89	0.93	0.98
	$\text{logit}^{-1}(\kappa_{ES}^S)$	0.23	0.26	0.28	0.31	0.35
	$\text{logit}^{-1}(\kappa_{ES}^R)$	0.67	0.73	0.77	0.81	0.86
	$\text{logit}^{-1}(\kappa_{FR})$	0.11	0.30	0.52	0.73	0.92
	$\text{logit}^{-1}(\kappa_{GR})$	0.11	0.31	0.55	0.76	0.93
	$\text{logit}^{-1}(\kappa_{HU})$	0.11	0.31	0.52	0.74	0.91
	$\text{logit}^{-1}(\kappa_{IE})$	0.81	0.89	0.93	0.96	0.98
	$\text{logit}^{-1}(\kappa_{IT}^S)$	0.34	0.38	0.40	0.43	0.48
	$\text{logit}^{-1}(\kappa_{IT}^R)$	0.37	0.43	0.46	0.49	0.54
	$\text{logit}^{-1}(\kappa_{LI})$	0.10	0.33	0.54	0.75	0.92
	$\text{logit}^{-1}(\kappa_{LT})$	0.58	0.64	0.69	0.73	0.81
	$\text{logit}^{-1}(\kappa_{LU})$	0.10	0.11	0.12	0.13	0.14
	$\text{logit}^{-1}(\kappa_{LV})$	0.37	0.41	0.44	0.47	0.52
	$\text{logit}^{-1}(\kappa_{MT})$	0.19	0.40	0.60	0.78	0.93
	$\text{logit}^{-1}(\kappa_{PL})$	0.26	0.29	0.31	0.33	0.37
	$\text{logit}^{-1}(\kappa_{PT})$	0.10	0.33	0.57	0.75	0.91
	$\text{logit}^{-1}(\kappa_{RO}^S)$	0.20	0.23	0.25	0.28	0.31
	$\text{logit}^{-1}(\kappa_{RO}^R)$	0.31	0.49	0.62	0.72	0.86
	$\text{logit}^{-1}(\kappa_{SI})$	0.59	0.66	0.70	0.76	0.85
	$\text{logit}^{-1}(\kappa_{SK})$	0.27	0.30	0.33	0.35	0.40
	$\text{logit}^{-1}(\kappa_{UK})$	0.41	0.45	0.47	0.50	0.54
Accuracy	$\tau_1^S$	1.2	33.3	616.9	910.3	1440.0
	$\tau_2^S$	16.7	18.1	19.2	20.4	22.3
	$\tau_3^S$	0.74	0.77	0.80	0.82	0.85
	$\tau_1^R$	45.7	80.5	142.7	303.9	820.1
	$\tau_2^R$	17.6	19.2	20.7	22.3	25.3
	$\tau_3^R$	1.16	1.22	1.27	1.32	1.41

Table 3: Posterior characteristics of the migration model parameters

	parameter	q5%	q25%	median	q75%	q95%
Intra-EU migration model	$\alpha_1$	6.89	7.04	7.12	7.20	7.31
	$\alpha_2$	0.27	0.30	0.33	0.36	0.39
	$\alpha_3$	0.23	0.26	0.29	0.32	0.36
	$\alpha_4$	-0.38	-0.19	-0.08	0.04	0.22
	$\alpha_5$	0.08	0.10	0.11	0.12	0.14
	$\alpha_6$	-0.40	-0.37	-0.34	-0.31	-0.27
	$\alpha_7$	0.00	0.08	0.14	0.20	0.27
	$\alpha_8$	0.17	0.24	0.29	0.35	0.42
	$\alpha_9$	-0.15	-0.08	-0.02	0.03	0.11
	$\alpha_{10}$	0.26	0.28	0.29	0.30	0.32
	$\alpha_{11}$	0.15	0.17	0.18	0.20	0.22
	$\alpha_{12}$	-0.16	-0.13	-0.11	-0.09	-0.07
	$\alpha_{13}$	-0.17	-0.14	-0.13	-0.11	-0.09
	$\alpha_{14}$	-0.21	-0.19	-0.17	-0.16	-0.14
	$\alpha_{15}$	-0.19	-0.17	-0.15	-0.14	-0.12
	$\alpha_{16}$	-0.12	-0.10	-0.09	-0.07	-0.05
	$\alpha_{17}$	-0.03	-0.01	0.00	0.01	0.03
	$\alpha_{18}$	0.39	0.53	0.66	0.77	0.94
	$\alpha_{19}$	0.19	0.22	0.24	0.26	0.29
	$\alpha_{20}$	0.91	1.08	1.17	1.29	1.45
Rest of world migration model	$\beta_1$	9.77	10.05	10.22	10.37	10.59
	$\beta_2$	0.33	0.54	0.68	0.81	0.99
	$\beta_3$	-0.21	0.16	0.39	0.59	0.97
	$\beta_4$	0.05	0.40	0.69	0.97	1.37
	$\beta_5$	0.01	0.22	0.35	0.49	0.67
	$\beta_6$	-2.63	-1.78	-1.24	-0.67	0.22
	$\beta_7$	-3.54	-1.13	1.04	2.83	5.86
	$\beta_8$	10.29	10.59	10.80	10.99	11.29
	$\beta_9$	0.70	0.96	1.11	1.29	1.53
	$\beta_{10}$	0.06	0.34	0.56	0.80	1.12
	$\beta_{11}$	-0.74	-0.29	0.00	0.33	0.81
	$\beta_{12}$	-0.49	-0.23	-0.04	0.14	0.41
	$\beta_{13}$	-3.06	-2.17	-1.56	-0.95	-0.16
	$\beta_{14}$	-2.91	-0.26	1.83	3.68	6.28
Precision	$\tau_y$	30.4	33.5	36.7	39.9	46.5
	$\tau_{0S}$	26.5	36.3	45.0	57.0	77.6
	$\tau_{0R}$	25.8	35.2	45.5	55.9	74.7

The accuracy of the data collecting systems is quantified by the precision of the error terms in the measurement equations. In Table 2, we present the posterior characteristics. Two important aspects of the data are worth noting. First, the measurement of immigration ( $\tau_a^R$ ) is more accurate than emigration ( $\tau_a^S$ ). Second, the most accurate are the Nordic countries ( $\tau_1^k$ ), followed by good registers ( $\tau_2^k$ ) and less reliable registers and surveys ( $\tau_3^k$ ).

For the undercount parameters, the estimated posterior medians remain close to the prior distributions specified in the model, i.e., 0.77 for ‘low undercount’ emigration, 0.86 for ‘low undercount’ immigration, 0.38 for ‘high undercount’ emigration and 0.70 for ‘high undercount’ immigration (see Table 2). As the identification of the undercount parameters is not possible from the data alone, the expert-based prior distributions were particularly informative for this part of our model.

We observe large differences in the posterior characteristics of the country-specific random effects that represent coverage (see Table 2). For example, according to the model, Bulgaria and Luxembourg are the most deficient in capturing migrants (median coverages are 7% and 12% respectively). In other words, having accounted for the general undercount and duration, the true flows to and from these countries are 14 and eight times larger than the reported figures, respectively. Apart from the Nordic countries and The Netherlands, for which the coverage was assumed to be perfect, high values of estimated coverage were achieved by the data collection systems of Germany, Ireland and Estonia. The precision of the random effects is relatively small for countries that provide no data (e.g., Belgium or Greece).

The integrated model for European migration produces posterior distributions for all true flows amongst the 31 countries from 2002 to 2008. Refer to the supplementary on-line materials for the full table of estimated median flows for the year 2008. For example, our median net migration totals presented in Figure 4 (solid line) imply that the overall gain in migration from the rest of world is around 820 thousand persons for 2008. Similar net migration totals were produced by the MIMOSA project (Raymer et al. 2011) and with our application of Abel’s (2010) approach (refer to Section 4.2). The corresponding figure resulting from adding up the published Eurostat data is around 1.5 million. Eurostat’s official figure, however, is likely to be overstated because it erroneously implies positive net migration within the (closed) EU and EFTA system.

Thus, there is a double-counting of migrants in official population totals caused by the different duration of migration measures used and the general underreporting of emigration found in the official statistics. Our approach models the full matrix which ensures a zero net migration within the EU and EFTA system.

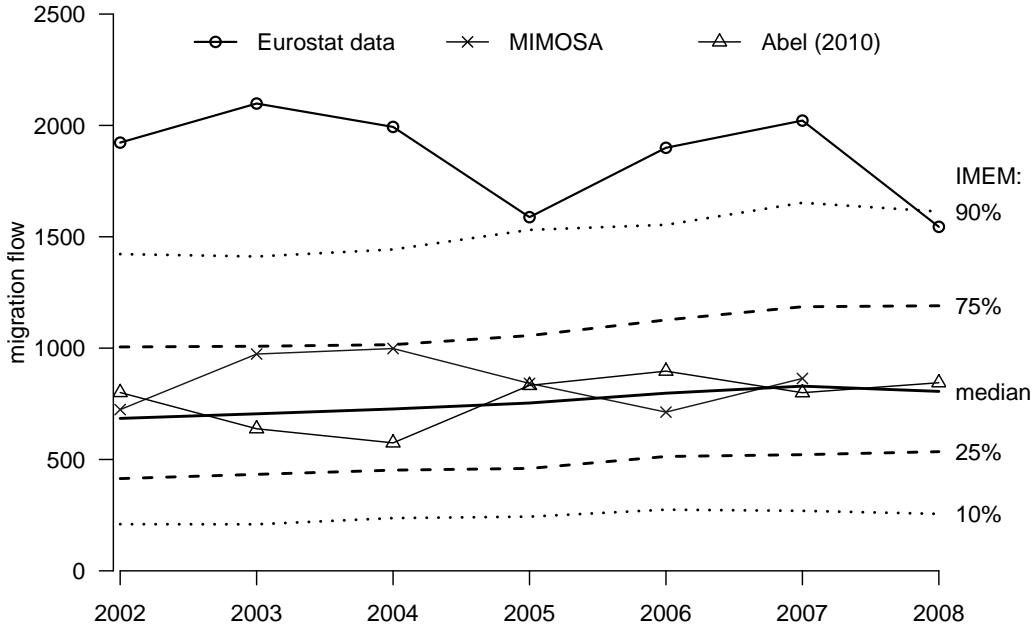


Figure 4: Reported and estimated net migration (in thousands) to EU and EFTA countries, 2002-2008

In Table 4, we present a subset of the 2008 median estimates for flows between countries with population sizes larger than 20 million. The corresponding median estimates for 2002-2008 are presented graphically in Figure 5, where for each flow, the scale on the vertical axis ranges from zero to twice the 2008 origin-destination median estimates. Table 4 and Figure 5 are meant to be used together. For example, for flow from Poland to the United Kingdom in 2008, the posterior median was estimated to be 83 thousand. We see from the patterns in Figure 5 that the levels increased considerably after 2004, and that they resemble the United Kingdom's reported statistics but not the Polish ones. In some cases, we only have the receiving country's report (e.g., France to the United Kingdom), whilst in other cases, we only have the sending country's report (e.g., United Kingdom to France).

In Figure 6, we present the posterior characteristics and densities of the 2006 flows from Finland to Denmark, from Denmark to The Netherlands, from the Czech Republic to Ireland, and from France to Hungary. For the Denmark to The Netherlands flow, both countries provided data, resulting in a posterior that is comparatively tight (the

Table 4: Median estimates of selected origin-destination flows in 2008

Origin	Destination						
	DE	ES	FR	IT	PL	RO	UK
DE		13330	16560	23380	104900	27240	15510
ES	12140		15730	5015	2756	17560	29270
FR	14480	10700		8321	6209	4978	49770
IT	17630	9190	14580		5484	14540	15210
PL	110200	7100	12640	16240		186	83020
RO	31000	64970	10750	72060	248		3062
UK	10800	27860	55660	9577	12370	918	

third-quartile-to-median ratio is 1.1). For the flow from France to Hungary, on the other hand, neither country provided data. Here, the posterior distribution is based primarily on the migration model. This flow is characterized by a relatively large amount of uncertainty and a heavy right tail (third-quartile-to-median ratio=1.9). The median flow from Finland to Denmark is characterized by relatively high precision (ratio=1.07), which results from the fact that these countries exchange their data on migrations. The last presented flow, from the Czech Republic to Ireland, is more uncertain with a median of 513 people and interquartile range of (400, 671) (ratio=1.3). Despite having both pieces of information about this flow, the Irish data are considered inaccurate due to the sampling error of the data source.

As another illustration, in Figure 7 we present the 2006 flows from Poland to Germany and from Finland to Sweden. The posterior true flow from Poland to Germany (top) has a median of 111900 people with interquartile range of (100400, 124300). Here, the reported data differ considerably from our estimated true flows. This is a consequence of 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 emigrants recorded (around 15 thousand). In the German data collection system, no time limit is applied for incoming flows.

In the bottom panel of Figure 7, the posterior density of the 2006 migration flow from Finland to Sweden is presented. The median is 3623 migrants with interquartile range of (3409, 3817). We also observe that the data reported by both sending and receiving countries are very close to each other (around 3100). Both reported flows lie in a tail of the posterior density and they are considerably lower than the median of the

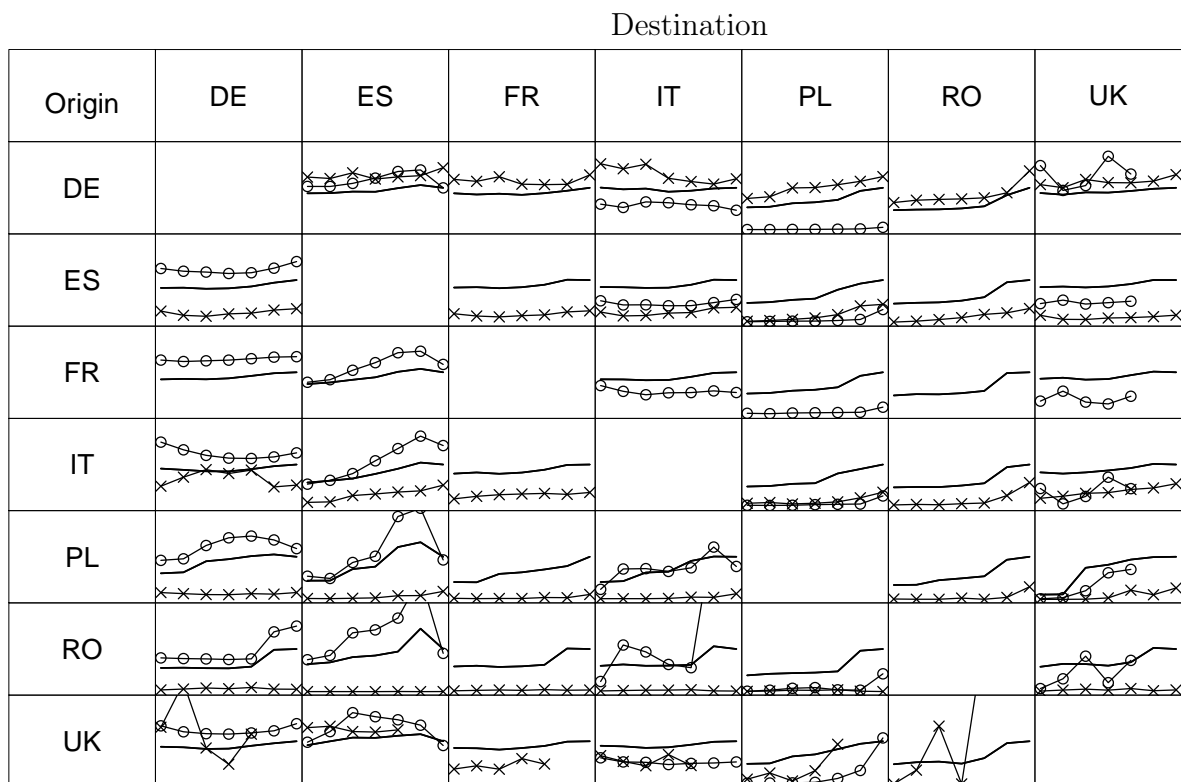


Figure 5: Median estimates of the selected true flows (solid), reported emigration (cross) and immigration (circle) data, 2002-2008

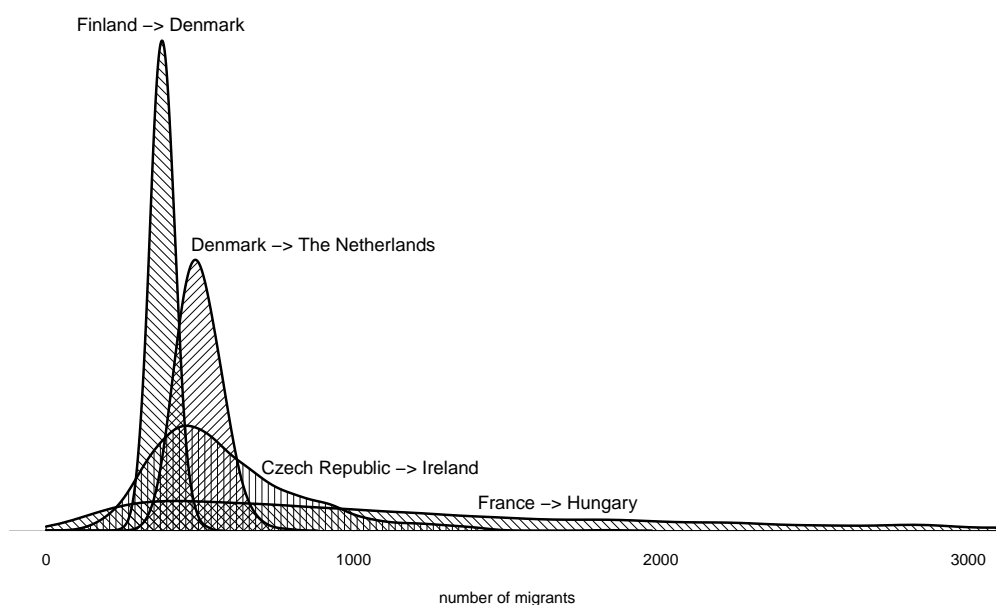


Figure 6: Posterior densities of the selected true migration flows, 2006

posterior true flow. This is due to our inclusion of expert information on the undercount of immigration and emigration and a very high precision of the estimate (the Nordic countries exchange information about the migration statistics). In De Beer et al. (2010)



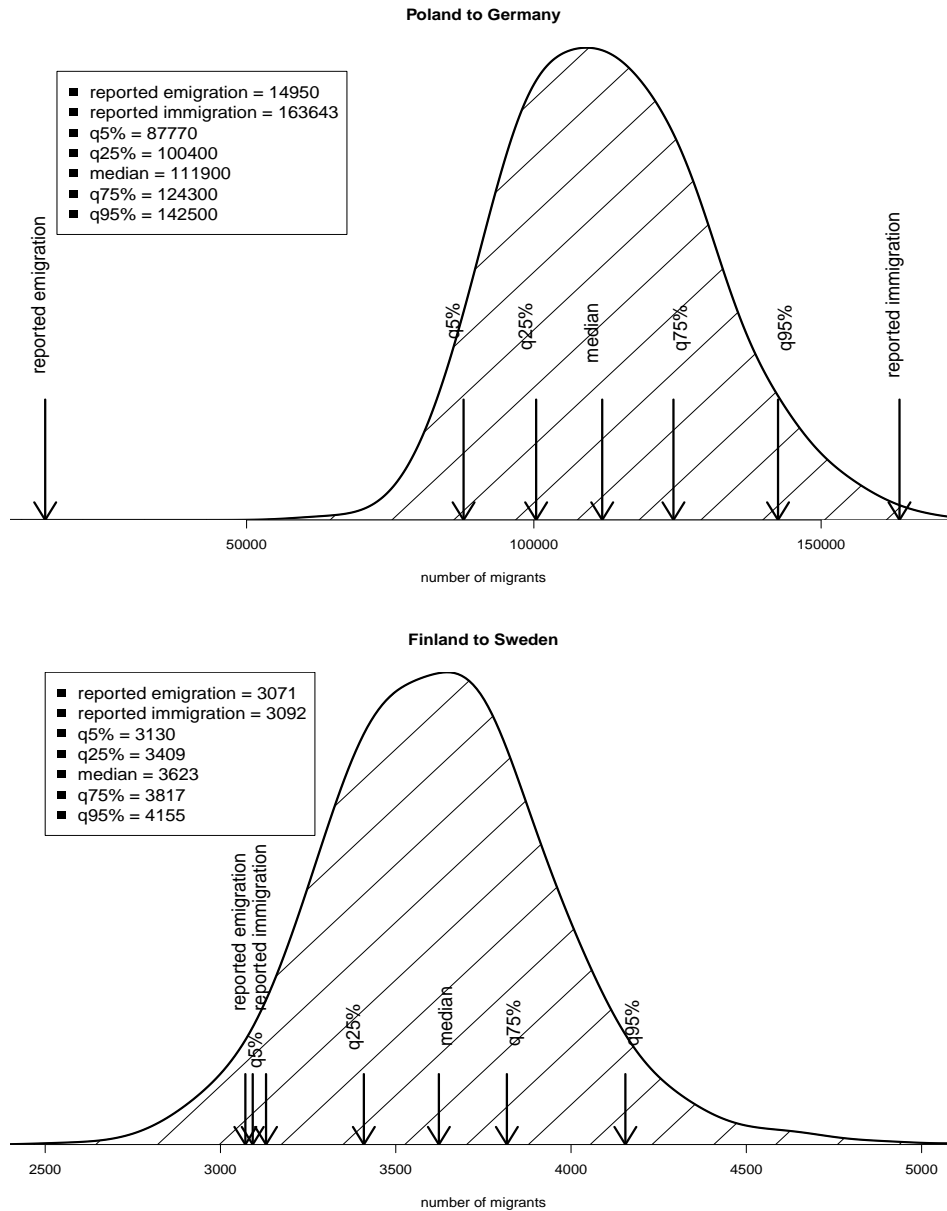


Figure 7: Posterior densities of migration flows from Poland to Germany (top) and Finland to Sweden (bottom) in 2006

and Raymer et al. (2011b), Sweden’s immigration data represented the benchmark and was assumed to be measured without error or undercount. In our model, the subjective expert assessment of the immigration undercount by means of prior distributions for  $\lambda_2$  and  $\lambda_4$  is incorporated. This leads to higher median flows than reported by the receiving countries, including the Nordic countries.

In Figures 8-11, the medians, 25th and 75th percentiles of the estimated total immigration, total emigration and net migration are presented for Sweden, the United Kingdom, Poland and France, respectively, along with the corresponding reported flows for comparison. Here, we see that immigration and emigration totals for Sweden are

slightly higher due to the inclusion of undercount in the measurement model. The interquartile ranges around these medians are narrow, as would be expected given the general high quality of data. For the UK, the interquartile ranges are much wider given its survey-based system. Our median estimates are, again, higher but they do not result in very different net migration totals. For Poland, the reported statistics are clearly too low and do not reflect the EU expansion in 2004. Finally, we present our estimates for France, a country that provided no origin-destination flows to Eurostat. They do, however, provide information on the total number of foreigners entering and leaving the country. While our emigration and immigration totals are higher than these reported figures, the median net migration comes close to the reported figures, albeit with a large amount of uncertainty.

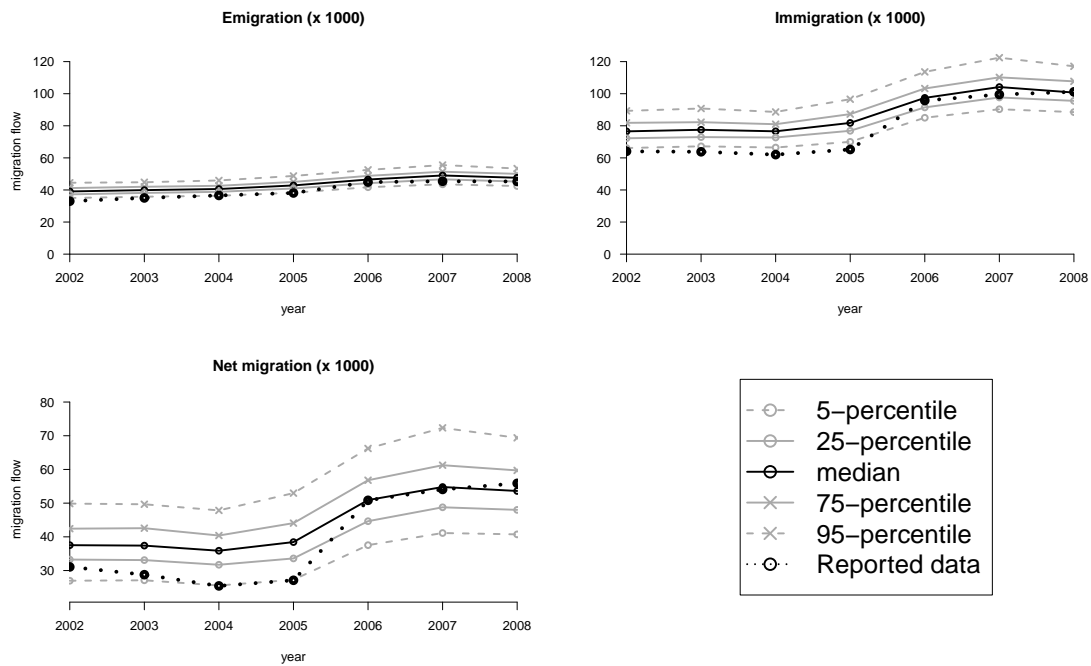


Figure 8: Estimated migration flows for Sweden

## 4.2 Model assessment

To assess the quality of the model, we first investigate the sensitivity of the results to changes in the assumptions regarding the prior densities for the parameters of the measurement model. Second, we analyse the sensitivity to the removal of 2008 data and some country-specific flows. Finally, we compare our results to other approaches

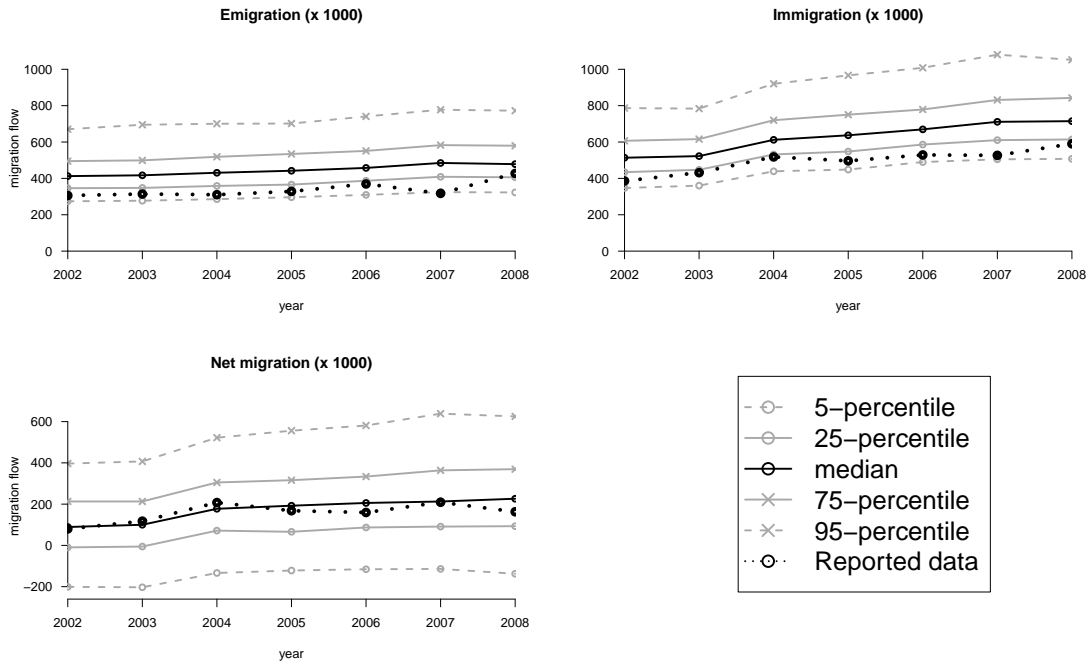


Figure 9: Estimated migration flows for the United Kingdom

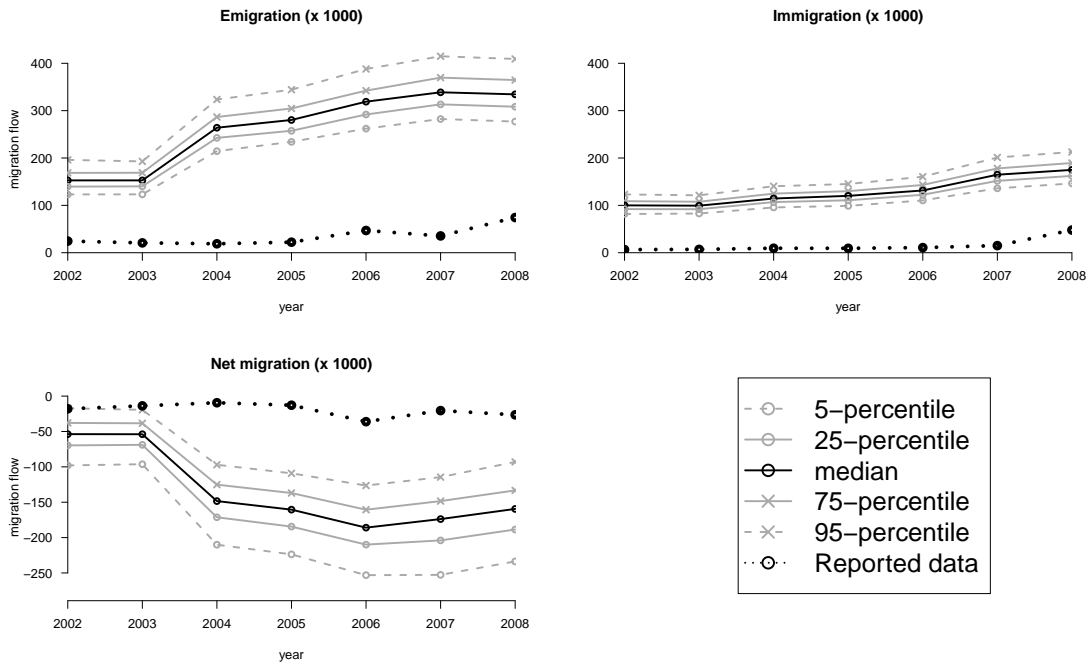


Figure 10: Estimated migration flows for Poland

that have been developed to estimate international migration flows (Abel, 2010; De Beer et al. 2010; Raymer et al. 2011b).

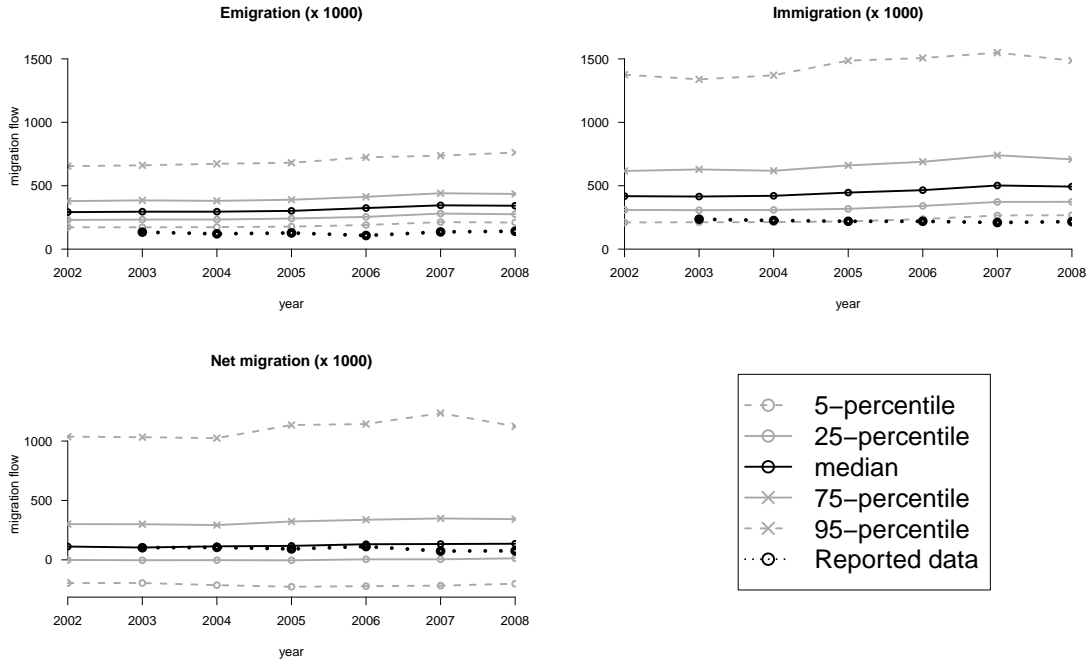


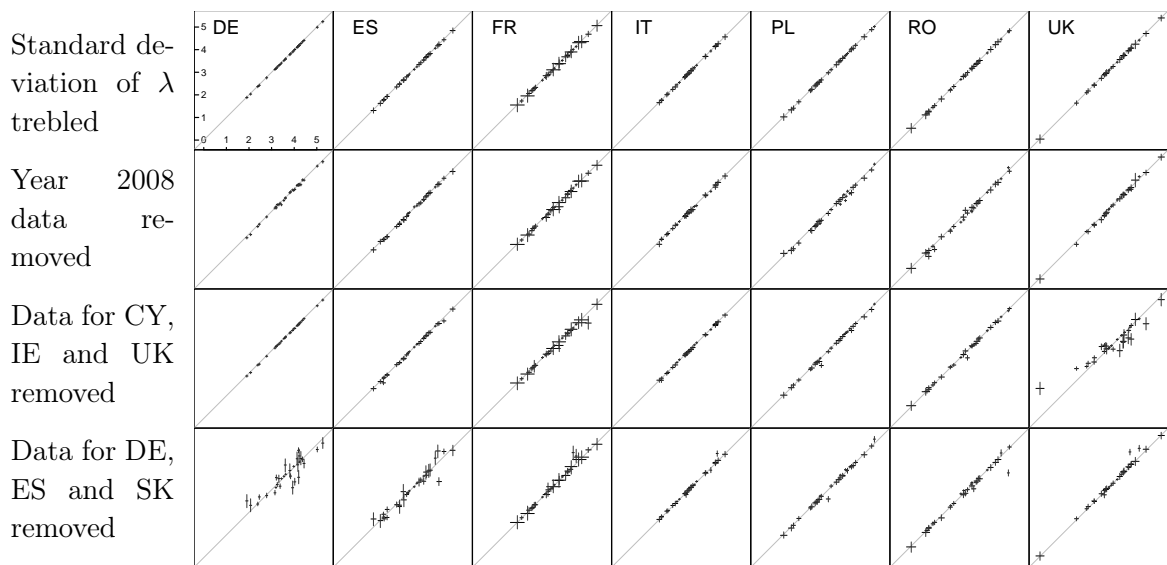
Figure 11: Estimated migration flows for France

#### 4.2.1 Sensitivity to the prior information

To analyze sensitivity to the prior for the undercount parameter, we kept the mean of the beta densities as elicited but doubled and trebled their standard deviations. For the auxiliary duration parameters  $d_m$ , we assumed weakly informative log-normal densities with mean zero and precision 0.05. To analyse the sensitivity of the accuracy parameters, we kept the classification of countries unchanged and applied the weakly informative gamma distributions  $\Gamma(0.1, 0.1)$  for the precisions of the error terms in the measurement model equations.

Doubling and trebling the standard deviations of the prior densities for the undercount parameters resulted in standard deviations of the posterior densities being 1.45-1.6 and 1.9-2.6 times larger, respectively (i.e., the increase in uncertainty of the parameters was less than proportional). The estimates of the true flows were even less sensitive to the increase in uncertainty of  $\lambda$ . In the first row of Figure 12, we present the medians and interquartile ranges (center and lengths of the cross, respectively) for flows among the seven largest countries for the original model and for the model with the trebled prior standard deviation for  $\lambda$ . The interquartile range for a given flow from the original model is horizontal and it cuts the corresponding vertical interquartile range for the flow from the validation model in its posterior median. Analogously, the

vertical interquartile range marks the posterior median of the flow from the original model. Hence, if the cross made by the two interquartile ranges lies on the  $45^\circ$  line, the medians are similar; if it lies below the line, the median from the validation model is smaller than median from the original one. In the first row of Figure 12, we do not observe large differences either in medians or the interquartile ranges. For the rest of the flows (not shown), the changes in medians rarely exceed 10% comparing to the original model and most of the differences regard the flows among countries with no data.



Note: The interquartile ranges for the original model are horizontal; the vertical interquartile ranges are for the validation models. The interquartile ranges of the final and validation models intersect at each other's medians. If the cross lies below the  $45^\circ$  line, the median of the flow from the validation model is lower than the one from the final model.

Figure 12: Comparison of interquartile ranges from the final model and models used for sensitivity analysis for true immigration flows (decimal logarithms) in 2008.

For all 6944 flows, we also computed the relative differences between the third-quarter-to-median ratios from the models with doubled and trebled standard deviation and the original one. For the model with the trebled prior standard deviation, the ratios that increased by more than five per cent constitute 3.4% of all ratios; the ratios that decreased by the same amount equaled 1.8% of all ratios. The share of ratios that changed by more than 10 per cent is 0.3%. We conclude that the results seem to be relatively insensitive to the inflation of uncertainty of the prior densities for the undercount parameters.

For the duration parameters, the uncertainty of the true flows is insensitive to the changes in the prior densities. In comparison to the original model, most of the differences in medians of the true flows based on the weakly informative prior densities oscillate around  $\pm 4\text{-}5\%$ , with the largest discrepancies of  $\pm 10\text{-}15\%$  concerning flows from and to the rest of world. Similarly, the results of the sensitivity analysis with respect to the precision of the error terms in the measurement model suggest that the posterior densities of the true flows remain mostly unaffected. The differences in medians are usually smaller than 5% in comparison with the original model. The same conclusion can be drawn about the posterior precision of the estimates.

#### 4.2.2 Sensitivity to the partial removal of the data

Sensitivity to the data can be assessed by removing subsets across time or across countries. Below, we describe the effects of removing (i) 2008 data, (ii) a selected group of countries with survey-based data collection systems (Cyprus, Ireland and the United Kingdom) and (iii) three sets of three randomly selected countries.

In the second, third and fourth rows of Figure 12, we compare the medians and interquartile ranges of the true flows from the models with 2008 data removed, survey-based countries removed and randomly selected countries (Germany, Spain and Slovakia) removed, respectively, with the corresponding characteristics from the original model. The largest sensitivity is observed for the flows based on the datasets with whole countries data removed (third and fourth rows). Removal of one year of the data does not affect the true flow estimates considerably. Not surprisingly, the largest differences in medians and uncertainty are observed for the countries where data have been removed. In the third row, it is the United Kingdom, and in the fourth row, it is Germany and Spain.

The quality of the model can be also evaluated by assessing how it predicts the data that have been removed. The predictive distributions of the data,  $\hat{z}_{ijt}^k$ , were computed as

$$p(\hat{z}|z) = \int p(\hat{z}|\theta)p(\theta|z)d\theta, \quad (11)$$

where  $z$  excludes 2008. Specifically, a sample of size 1000 from the posterior distribution of each of the parameters was saved. Then, for each  $n$ ,  $n = 1, \dots, 1000$ , the  $\log \bar{\mu}_{ijt}^k(n)$

were computed as

$$\log \bar{\mu}_{ijt}^S(n) = \log y_{ijt}(n) + \delta_{m(i)}(n) - \log \lambda_{f(i)}(n) - \log (1 + e^{-\kappa_i(n)}), \quad (12)$$

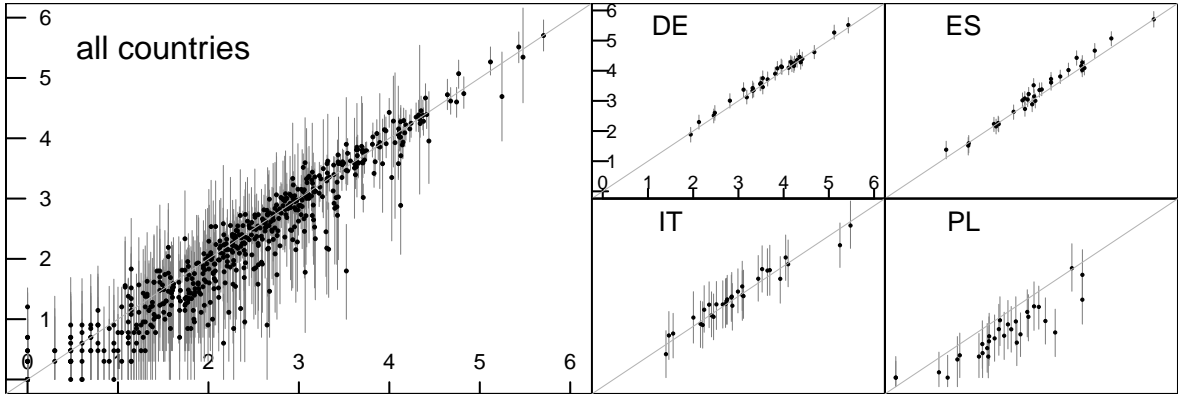
$$\log \bar{\mu}_{ijt}^R(n) = \log y_{ijt}(n) + \delta_{m(j)}(n) - \log \lambda_{g(j)}(n) - \log (1 + e^{-\kappa_j(n)}). \quad (13)$$

For the true flows  $y_{ijt}(n)$ , we used samples from the migration model (Equations 8, 9 and 10) for parameters  $\alpha(n)$ ,  $\beta(n)$ ,  $u_{ij}(n)$ ,  $\tau_y(n)$ ,  $\tau_{0R}(n)$  and  $\tau_{0S}(n)$ . In the next step,  $\log \mu_{ijt}^k(n)$  were drawn from normal distributions:  $\log \mu_{ijt}^k(n) \sim \mathcal{N}(\log \bar{\mu}_{ijt}^k(n), \varepsilon_{ijt}^k(n))$ . Then, 1000 replications of the predicted data were drawn:  $\hat{z}_{ijt}^k(n) \sim \text{Po}(\mu_{ijt}^k(n))$ .

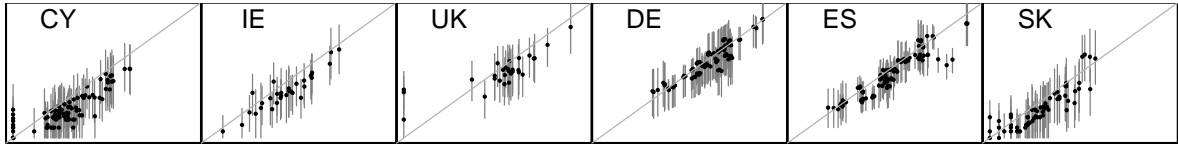
In Figure 13, we present the medians (dots) and 95% predictive intervals (vertical lines) for the predictive distributions of the data  $\hat{z}$  (decimal logarithms) for immigration. For the model with 2008 data removed, the characteristics of predictive distributions in that year for all countries are shown in Panel A. In Panel B, the same characteristics are replotted for the four largest countries with available data. The predictive intervals cover the 45° line in majority of the cases. For countries with reliable immigration registers (Germany, Spain), the uncertainty spans are narrower than for those with less reliable data (Poland). Also, for Poland, the predicted medians are often smaller than the reported data. In general, we find that 54% of observed emigration and 43% of observed immigration in 2008 fall into the interquartile ranges of the predictive distributions for  $\hat{z}$ . For the 90% predictive intervals, the respective shares are 86% and 82%. The shares of observed data that are larger than the medians for emigration and immigration are 58% and 65%, respectively, whereas the shares of the data smaller than medians are 47% and 40%. These results suggest that the model tends to underpredict the data.

For the models with the data removed for specific countries, in Panels C-F in Figure 13, we present the 95% predictive intervals for the data for these countries. We observe that the 95% predictive intervals usually cover the 45° line. For the model with survey-based data countries removed, the percentages of the observations within and above the the 90% predictive interval are 72% and 18%, respectively. For the models with three countries randomly removed, we obtained the following percentages of the observations within and above the the 90% predictive interval: 84% and 16%, respectively, when Germany, Spain and Slovakia were removed; 95% and 1%, respectively, when Italy,

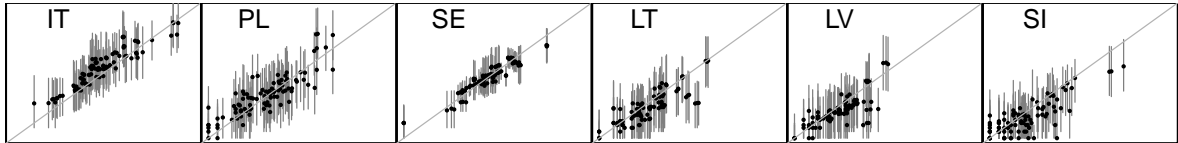
A. Model with data for 2008 removed; all countries in 2008 presented      B. Model with data for 2008 removed; example countries in 2008 presented



C. Model with data for CY, IE and UK removed; 2006-2008 presented      D. Model with data for DE, ES and SK removed; 2006-2008 presented



E. Model with data for IT, PL and SE removed; 2006-2008 presented      F. Model with data for LT, LV and SI removed; 2006-2008 presented



Note: Data  $z$  on horizontal axis, medians (dots) and 95% predictive intervals (vertical dark-grey lines) of the predictive distributions for  $\hat{z}$  on vertical axis.

Figure 13: Comparison of the data  $z$  on immigration flows and predictive distributions of the data,  $\hat{z}$  (decimal logarithms).

Poland and Sweden were removed; and 94% and 6%, respectively, when Lithuania, Latvia and the Slovenia were removed. For several countries, such as Cyprus, Ireland, United Kingdom, Spain, Slovakia, Lithuania or Latvia, the posterior medians are often lower than the officially reported figures, which indicates that the model underpredicts the data for these countries. Overprediction is clearly only observed for Italy.

#### 4.2.3 Comparison with the alternative approaches

To assess the quality of the model, we have compared the estimated flows with the results obtained by applying the method developed by Abel (2010) to the set of data on intra-European flows only and the set of covariates as in our model (except for the random effects) and from the MIMOSA project (De Beer et al. 2010; Raymer et al.



2011b). For Abel’s method, we first obtained a set of correction factors from a constrained optimization procedure applied to the available sending and receiving country data that were deemed to be reliable. The baseline flows represented immigration flows for all Nordic countries and The Netherlands plus Sweden’s emigration flows, i.e., the correction factors for these countries were set to one. For the procedure, the Canberra distance measure was chosen as recommended by Abel (2010: 811). Next, all reported flows were harmonized by using these correction factors. Finally, the completely missing data were estimated by using a negative binomial regression fitted to the harmonized data. In the MIMOSA project, a similar approach was applied but for the computation of the correction factors. Here, additional expert knowledge was incorporated by means of indicator variables assigned to selected flows (see De Beer et al. 2010). The explanatory model for missing data was a normal regression with a different set of covariates (see Raymer et al. 2011b).

The reported and estimated median flows amongst the seven largest countries in the EU and EFTA are presented in Table 5, along with estimates from the MIMOSA project and from the application of Abel’s method to our data and covariates. We observe that, with the exception of flows from Germany to Italy and Romania, all reported data fall outside our 50% predictive intervals, and most are outside the 95% predictive intervals. The MIMOSA and Abel estimates are more in line with our median estimates, with most falling within our 95% predictive intervals. However, it is not clear which of the two methods produces more similar results to ours. For example, Abel’s method produces results that are close to ours for migration from Germany, whereas from Spain, the MIMOSA estimates were closer. The largest discrepancies concern the flows between Poland and the United Kingdom and between Italy and Romania. As mentioned in Section 2, our approach to harmonizing and estimating migration differs from MIMOSA and Abel by its emphasis on modeling the measurement aspects of the reported statistics and by providing measures of uncertainty for all flows and parameters in the model.

Table 5: Comparison of estimates for selected origin-destination flows in 2007

Origin	Destination						
	DE	ES	FR	IT	PL	RO	UK
DE	Emigr.	17123 <sup>‡</sup>	17912 <sup>‡</sup>	25413 <sup>‡</sup>	120791 <sup>‡</sup>	24054 <sup>‡</sup>	17942 <sup>‡</sup>
	Immigr.	18902 <sup>‡</sup>	NA	13562 <sup>‡</sup>	3913 <sup>‡</sup>	NA	NA
	Abel	14052 <sup>‡</sup>	14709 <sup>‡</sup>	18360 <sup>‡</sup>	59575 <sup>‡</sup>	19753 <sup>‡</sup>	14734 <sup>‡</sup>
	MIMOSA	15343 <sup>‡</sup>	12704 <sup>‡</sup>	18920 <sup>‡</sup>	110379 <sup>‡</sup>	17061 <sup>‡</sup>	22731 <sup>‡</sup>
	IMEM	14180	15420	23000	97330	22560	15150
ES	Emigr.	3601 <sup>‡</sup>	3986 <sup>‡</sup>	1740 <sup>‡</sup>	1087 <sup>‡</sup>	4078 <sup>‡</sup>	4161 <sup>‡</sup>
	Immigr.	15515 <sup>‡</sup>	NA	2366 <sup>‡</sup>	213 <sup>‡</sup>	NA	NA
	Abel	16290 <sup>‡</sup>	14132 <sup>‡</sup>	3203 <sup>‡</sup>	3243 <sup>‡</sup>	14459 <sup>‡</sup>	14753 <sup>‡</sup>
	MIMOSA	12543 <sup>‡</sup>	28803 <sup>‡</sup>	3020 <sup>‡</sup>	2622 <sup>‡</sup>	19567 <sup>‡</sup>	17408 <sup>‡</sup>
	IMEM	11380	15850	5054	2526	16640	29300
FR	Emigr.	NA	NA	NA	NA	NA	NA
	Immigr.	19627 <sup>‡</sup>	15973 <sup>‡</sup>	4696 <sup>‡</sup>	394 <sup>‡</sup>	NA	NA
	Abel	20608 <sup>‡</sup>	11874 <sup>‡</sup>	6357 <sup>‡</sup>	5999 <sup>‡</sup>	5677 <sup>‡</sup>	25932 <sup>‡</sup>
	MIMOSA	15868 <sup>‡</sup>	11671 <sup>‡</sup>	6335 <sup>‡</sup>	6618 <sup>‡</sup>	2036 <sup>‡</sup>	22573 <sup>‡</sup>
	IMEM	14160	11550	8191	5694	4885	50560
IT	Emigr.	8382 <sup>‡</sup>	3520 <sup>‡</sup>	4395 <sup>‡</sup>	1211 <sup>‡</sup>	3963 <sup>‡</sup>	6870 <sup>‡</sup>
	Immigr.	20771 <sup>‡</sup>	15301 <sup>‡</sup>	NA	412 <sup>‡</sup>	NA	NA
	Abel	21809 <sup>‡</sup>	11375 <sup>‡</sup>	14203 <sup>‡</sup>	6273 <sup>‡</sup>	12807 <sup>‡</sup>	22202 <sup>‡</sup>
	MIMOSA	16793 <sup>‡</sup>	8947 <sup>‡</sup>	7524 <sup>‡</sup>	6794 <sup>‡</sup>	1818 <sup>‡</sup>	6960 <sup>‡</sup>
	IMEM	16950	9608	14440	4904	13640	15430
PL	Emigr.	13771 <sup>‡</sup>	650 <sup>‡</sup>	533 <sup>‡</sup>	813 <sup>‡</sup>	9 <sup>‡</sup>	9165 <sup>‡</sup>
	Immigr.	153589 <sup>‡</sup>	15166 <sup>‡</sup>	NA	19936 <sup>‡</sup>	NA	NA
	Abel	161264 <sup>‡</sup>	11275 <sup>‡</sup>	10325 <sup>‡</sup>	26989 <sup>‡</sup>	174 <sup>‡</sup>	177546 <sup>‡</sup>
	MIMOSA	124171 <sup>‡</sup>	7946 <sup>‡</sup>	8836 <sup>‡</sup>	13361 <sup>‡</sup>	77 <sup>‡</sup>	43888 <sup>‡</sup>
	IMEM	116200	9514	9886	16330	173	82360
RO	Emigr.	1902 <sup>‡</sup>	138 <sup>‡</sup>	372 <sup>‡</sup>	1401 <sup>‡</sup>	3 <sup>‡</sup>	71 <sup>‡</sup>
	Immigr.	43456 <sup>‡</sup>	174217 <sup>‡</sup>	NA	274007 <sup>‡</sup>	9 <sup>‡</sup>	NA
	Abel	45627 <sup>‡</sup>	129514 <sup>‡</sup>	16281 <sup>‡</sup>	370942 <sup>‡</sup>	137 <sup>‡</sup>	3107 <sup>‡</sup>
	MIMOSA	35133 <sup>‡</sup>	91820 <sup>‡</sup>	5674 <sup>‡</sup>	74512 <sup>‡</sup>	279 <sup>‡</sup>	1722 <sup>‡</sup>
	IMEM	30430	95910	10880	76690	238	3145
UK	Emigr.	NA	NA	NA	NA	NA	NA
	Immigr.	13443 <sup>‡</sup>	38347 <sup>‡</sup>	NA	4703 <sup>‡</sup>	3913 <sup>‡</sup>	NA
	Abel	14115 <sup>‡</sup>	28507 <sup>‡</sup>	23958 <sup>‡</sup>	6367 <sup>‡</sup>	59575 <sup>‡</sup>	1681 <sup>‡</sup>
	MIMOSA	10868 <sup>‡</sup>	37128 <sup>‡</sup>	20915 <sup>‡</sup>	6725 <sup>‡</sup>	9117 <sup>‡</sup>	719 <sup>‡</sup>
	IMEM	10290	32470	55130	9457	11630	877

Notes: (i) Emigr. refers to reported emigration flows; (ii) Immigr. refers to reported immigration flows; (iii) Abel refers to the application of the method described in Abel (2010) using our data and covariates; (iv) MIMOSA refers to the estimates described in Raymer et al. (2011b); and (v) IMEM refers to the median estimates of flows from the model as described in this paper. The symbols ‡, † and ‡ in superscripts denote that the number falls into, respectively, 50%, 95% and outside the 95% predictive interval of the posterior density of a given IMEM flow.

## 5 CONCLUSIONS

There are four main contributions of this research. First, we have created a methodology for estimating international migration flows that directly accounts for the main differences found in the measurement aspects of the reported data. Different measurements of migration flows can result in very different patterns, as we have seen with figures reported by Germany and Poland. However, before this study, little was known about the effects of measurement and error, and no one had attempted to model the differences by considering the main aspects of duration, undercount, coverage and accuracy. We have done so by including expert-based prior distributions in our model.

Second, we have combined the measurement model with a migration model. This allowed us to bring together the reported data, covariate information and expert judgments. The estimated flows are consistent with the United Nations recommendation for the measurement of international migration. Third, our estimated flows can include measures of uncertainty, such as posterior predictive intervals, which can be used to assess the quality of the reported flows. When combining data from different sources and estimated missing data, it is important to be clear about the accuracy of the estimated figures.

Finally, we have produced a consistent and complete set of estimates that can be used by the wider community. This work is especially relevant considering the expansions of the European Union in 2004 and 2007. Our results can be used to inform policy and to improve the evidence base. For example, our estimated net migration totals, in comparison to those published by Eurostat, suggest that the official population totals for all the EU and EFTA countries are likely to be about one million too high for the years 2002-2008.

With respect to making full use of the outcomes of the model presented in this paper, it is crucial to stress that whole posterior distributions of the estimates provide much more information than point estimates alone. Demographic literature provides some suggestions on how this information can be subsequently utilized by the users of the estimates, such as policy makers or public planners. In this context, the main idea is to apply Bayesian decision theory. By taking into account potential losses involved in decision making, the decisions are optimized given the estimated uncertainty of the

quantities of interest. Alho and Spencer (2005) and Bijak (2010) provide an overview of decision theory applications in the context of population estimates. They note that loss functions may differ depending on the purpose of producing estimates, whether these are needed for the allocation of resources, or purely for research.

Importantly from the point of view of the work on European migration, Alho and Spencer (2005: 368) outline a cost-benefit analysis for data improvement programs, noting that this should at least focus attention on the relative importance of different elements of the data collection process. The posterior summaries presented in this paper correspond to specific loss functions (for example, the median estimates correspond to a situation with linear losses which are symmetric for over- and underestimation of migration flows). Bijak (2010) argues that given this specificity, further applied work in this area should require a closer collaboration between statisticians producing the estimates, and decision makers using their results, especially with respect to elicitation of such context-specific loss functions.

In the context of the model presented in this paper, more research needs to be undertaken for the rest of world component and for incorporating reported flows on the total immigration and emigration movements into the model. Migration data for countries in the rest of world are very limited. However, it is possible the estimates could improve by disaggregating this single category into more categories, for example referring to continents or other country groupings.

We also did not take into account some information provided by countries on the total flows of migrants. For example, information is available on the total immigration and emigration for Switzerland, a country that only collects data on the citizenship of migrants and not country of origin or destination of migrants. The model presented in this paper could be extended by building a separate set of measurement and migration models for the total inflows and outflows for all countries.

Another path of further study includes the elicitation and quantification of expert opinion. An alternative approach to the one presented in this paper could utilize detailed information obtained from experts on the country-specific data collection systems. This could permit the creation of more adequate typologies of undercount or accuracy. The crucial expert opinion is required on the magnitude of the immigration undercount, as no data permits identification of its value, which can vary across

countries, as well as over time.

To conclude, we have presented a modeling framework that brings together empirical data, covariate information and expert judgments to estimate migration flows amongst 31 countries in Europe from 2002 to 2008. This work provides an important foundation for both modeling 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 was 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. We would also like to thank the other IMEM team members: Guy J. Abel, Solveig Christiansen, Joop de Beer, Nico Keilman, Jeannette Schoorl and Rob van der Erf, as well as the Editor, the Associate Editor and two reviewers, for their valuable comments and suggestions on this work.

## References

- Abel, G. J. (2010), "Estimation of International Migration Flow Tables in Europe," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 173(4), 797-825.
- Abel, G. (2012). "Estimating Global Migration Flow Tables Using Place of Birth Data," Vienna Institute of Demography Working Paper 01/2012.
- Alho, J. M. and Spencer, B. D. (2005), *Statistical Demography and Forecasting*, New York: Springer.
- Bijak, J. (2010), *Forecasting International Migration in Europe: A Bayesian View*, Dordrecht: Springer.

- 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.
- Bilsborrow, R. E., Hugo, G., Oberai A. S., and Zlotnik, H. (1997), *International Migration Statistics: Guidelines for Improving Data Collection Systems*, Geneva: International Labour Office.
- Brierley, M. J., Forster, J. J., McDonald, J. W., and Smith, P. W. F. (2008), “Bayesian Estimation of Migration Flows,” in *International Migration in Europe: Data, Models and Estimates*, eds. J. Raymer and F. Willekens, Chichester, Wiley, pp. 149-174.
- Castles, S. and Miller, M. J. (2009), *The Age of Migration: International Population Movements in the Modern World*, 4th Edition, Basingstoke: Palgrave Macmillan.
- Cohen, J. E., Roig, M., Reuman, D. C., 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.
- 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.
- DeWaard, J., Kim, K., and Raymer, J. (2012), Migration Systems in Europe: Evidence from Harmonized Flow Data, *Demography*, 49, 1307-1333.
- European Commission (2006), “FAQ on the Commission’s Free Movement of Workers Report,” MEMO/06/64, European Commission, Brussels.
- Hill, K. (1985), “Indirect Approaches to Assessing Stocks and Flows of Migrants,” in *Immigration Statistics: A Story of Neglect*, eds. D. B. Levine, K. Hill, and R. Warren, Washington, DC: National Academy Press, pp. 205-224.
- Goldin, I., Cameron, G., and Balarajan, M. (2010), *Exceptional People: How Migration Shaped Our World and Will Define Our Future*, Princeton: Princeton University Press.

- Jasso, G. and Rosenzweig, M. R. (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.
- Jennissen, R. (2004), *Macro-Economic Determinants of International Migration in Europe*, PhD Thesis, Rijksuniversiteit Groningen.
- Kelly, J. J. (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.
- Kim, K. and Cohen, J. E. (2010), "Determinants of International Migration Flows To and From Industrialized Countries: A Panel Data Approach Beyond Gravity," *International Migration Review*, 44(4), 899-932.
- 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*, eds. J. Raymer and F. Willekens, Chichester: Wiley, pp. 41-71.
- 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.
- Lewis, M. (ed.), (2009), *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com/>.
- Mayer, T. and Zignago, S. (2006), "Notes on CEPIIs distances measures," Centre d'Etudes Prospectives d'Informations Internationales (CEPII), Paris.
- Nowok, B. (2010), *Harmonization by Simulation: A Contribution to Comparable International Migration Statistics in Europe*, Amsterdam: Rozenberg Publishers.
- Nowok, B., Kupiszewska, D., and Poulain, M. (2006), "Statistics on International Migration Flows, in *THESIM: Towards Harmonised European Statistics on International Migration*, eds. M. Poulain, N. Perrin, and A. Singleton, Louvain-la-Neuve: UCL Presses, pp. 203-231.

- Nowok, B. and Willekens, F. (2011), "A Probabilistic Framework for Harmonisation of Migration Statistics," *Population, Space and Place*, 17(5), 521-533.
- Özden, C., Parsons, C., Schiff, M., and Walmsley T. (2011). Where on earth is everybody? The evolution of global bilateral migration 1960-2000. *The World Bank Economic Review*, 25, 12-56.
- Parsons, C. R., Skeldon, R., Walmsley, T. L., and Winters, L. A. (2007), "Quantifying international migration: a database of bilateral migrant stocks," World Bank Policy Research Working Paper 4165, March 2007, The World Bank, Washington.
- Poulain, M. (1993), "Confrontation des Statistiques de Migrations Intra-Européennes: Vers Plus D'harmonisation?," *European Journal of Population*, 9, 353-381.
- Poulain, M., Perrin, N., and Singleton, A., eds. (2006), *THESIM: Towards Harmonised European Statistics on International Migration*, Louvain: UCL Presses.
- 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.
- Raymer, J. (2008), "Obtaining an overall picture of population movement in the European Union," in *International Migration in Europe: Data, Models and Estimates*, eds. J. Raymer and F. Willekens, Chichester: Wiley, pp. 209-234.
- Raymer, J., Abel, G. J., Disney, G., and Wiśniowski, A. (2011a), "Improving Estimates of Migration Flows to Eurostat," CPC Working Paper 15, ESRC Research Centre for Population Change, University of Southampton.
- Raymer, J., De Beer, J., and Van der Erf, R. (2011b), "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.
- Rowe, G. and Wright, G. (2001), "Expert Opinions in Forecasting: Role of the Delphi Technique," in *Principles of Forecasting: A Handbook of Researchers and Practitioners*, ed. J. S. Armstrong, Boston: Kluwer Academic Publishers, pp. 125-144.



- Spiegelhalter, D., Thomas, A., and Best, N. (2011), *OpenBUGS User Manual*, Version 3.2.1 [online], Available at <http://www.mrc-bsu.cam.ac.uk/bugs>.
- 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.
- Van der Erf, R. (2009), "Typology of Data and Feasibility study," MIMOSA Deliverable 9.1 B Report, Netherlands Interdisciplinary Demographic Institute, The Hague.
- Van der Erf, R. and Van der Gaag, N. (2007), "An Iterative Procedure to Revise Available Data in the Double Entry Matrix for 2002, 2003 and 2004," MIMOSA Discussion Paper, Netherlands Interdisciplinary Demographic Institute, The Hague.
- Van der Gaag, N. and Van Wissen, L. (2002), "Modelling Regional Immigration: Using Stocks to Predict Flows," *European Journal of Population*, 18, 387-409.
- Warren, R. and Peck, J. M. (1980), "Foreign-Born Emigration from the United States: 1960 to 1970," *Demography*, 17(1), 71-84.
- 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.
- World Bank (2010), "World Development Indicators" [online], The World Bank, Washington, Available at:  
<http://data.worldbank.org/data-catalog/world-development-indicators>.
- Zaba, B. (1987), "The Indirect Estimation of Migration: A Critical Review," *International Migration Review*, 21(4), 1395-1445.
- Zlotnik, H. (1987), "The Concept of International Migration as Reflected in Data Collection Systems," *International Migration Review*, 21(4), 925-946.

## Supplementary on-line material

Table A and Figure A are meant to be used together. In Table A, we present the 2008 median estimates of flows amongst 31 countries under study and the rest of the world. The corresponding median estimates for 2002-2008 are presented graphically in Figure A. For each of these flows, the scale on the vertical axis ranges from zero to double the 2008 origin-destination median estimates presented in Table A. Country codes: AT - Austria, BE - Belgium, BG - Bulgaria, CH - Switzerland, CY - Cyprus, CZ - Czech Republic, DE - Germany, DK - Denmark, EE - Estonia, ES - Spain, FI - Finland, FR - France, GR - Greece, HU - Hungary, IE - Ireland, IS - Iceland, IT - Italy, LI - Liechtenstein, LT - Lithuania, LU - Luxembourg, LV - Latvia, MT - Malta, NL - Netherlands, NO - Norway, PL - Poland, PT - Portugal, RO - Romania, SE - Sweden, SI - Slovenia, SK - Slovakia, UK - United Kingdom, RW - rest of the world.

For more information on Integrated Modelling of European Migration project and results, visit <http://www.imem.cpc.ac.uk>.

Table A: Median estimates of origin-destination flows in 2008

Origin	Destination															
	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HU	IE	IS
AT	230	1269	1363	26	1198	10730	197	29	691	189	722	397	2727	133	28	
BE	268	357	1391	57	235	3305	440	78	2666	241	22840	1303	349	440	23	
BG	2192	1363	496	1549	3489	14720	321	21	13950	140	2086	8812	134	181	13	
CH	1053	1109	112	42	257	8570	452	26	3123	277	6948	788	333	214	25	
CY	26	27	558	15	33	240	21	5	31	35	89	2357	92	33	1	
CZ	1379	398	2600	438	133	6635	240	31	778	90	1947	328	214	569	41	
DE	17170	4019	10570	19690	285	6908	3495	605	13330	1559	16560	14570	15780	2092	245	
DK	172	366	97	496	21	2364	119	119	998	496	1146	144	194	230	1161	
EE	49	108	14	42	30	577	183	12	142	3769	180	14	15	284	25	
ES	748	4680	3737	6285	62	12140	1426	91	753	870	15730	469	496	2324	62	
FI	180	294	26	421	35	67	302	1084	753	462	94	151	241	40		
FR	873	25480	692	8760	203	1100	1291	119	10700	468	2513	2513	1404	2585	56	
GR	367	1366	2737	907	4594	171	7565	12	482	115	2372	160	355	143	3	
HU	4164	695	84	834	202	17200	345	24	889	276	2571	597	176	908	21	
IE	138	508	93	233	62	1517	264	81	1462	212	2248	160	7	10	9	
IS	29	24	5	25	1	209	1189	15	120	64	65	4	7	10	9	
IT	1717	5216	1464	19310	60	683	828	103	9190	373	14580	1207	972	1262	62	
LI	43	5	1	175	0	1	9	0	15	1	33	7	2	1	0	
LT	174	198	62	124	30	138	1027	147	1684	126	583	77	18	3690	198	
LU	75	3969	16	342	3	15	159	15	124	69	5032	59	39	85	29	
LV	87	74	24	90	50	37	436	369	342	140	348	22	13	1058	64	
MT	17	38	14	33	34	9	33	1	27	10	182	31	7	52	3	
NL	835	15110	398	1736	83	554	721	52	4461	370	4984	875	708	813	80	
NO	74	107	32	132	23	43	2180	49	741	827	314	50	39	54	309	
PL	4905	4329	327	2282	496	3297	3441	50	7100	442	12640	2954	399	10650	1344	
PT	302	1567	250	2926	14	67	185	36	10570	86	28950	207	93	395	60	
RO	7550	1593	70	3517	1008	2431	688	20	64970	183	10750	2647	17310	811	30	
SE	357	533	95	810	113	169	3448	231	1569	4386	1396	689	272	361	435	
SI	663	134	21	498	6	56	50	2	144	16	265	26	43	44	9	
SK	3772	206	134	1231	127	26990	204	9	666	51	814	105	930	398	34	
UK	1236	5900	589	9094	5358	1952	2957	152	27860	1318	55660	12200	3831	18690	132	
RW	36550	30930	5019	27090	7378	53300	17650	2848	309100	12670	219800	23830	20640	24670	1020	
Total	87909	127614	33249	125166	22999	107058	44819	6540	488932	30054	493007	85927	73598	80530	5683	

Table A (continued): Median estimates of origin-destination flows in 2008

Origin	Destination																Total
	IT	LI	LT	LU	LV	MT	NL	NO	PL	PT	RO	SE	SI	SK	UK	ROW	
AT	1261	66	128	62	55	12	574	129	3623	233	4857	448	599	2570	1200	26620	62955
BE	2499	7	92	4173	49	21	6307	203	1887	982	685	569	114	210	6365	19630	87659
BG	6179	2	63	42	49	24	1843	244	664	526	131	642	45	484	4988	5376	73492
CH	9218	236	51	197	40	17	791	186	1026	1522	1166	692	400	704	7486	38480	94431
CY	52	0	10	2	15	9	65	18	132	4	487	110	3	48	2853	6090	13924
CZ	1233	1	87	30	58	10	726	145	1915	97	2580	347	65	27010	6509	62870	121105
DE	23380	80	2380	1618	1408	118	9514	2731	104900	6401	27240	4172	1783	8574	15510	184100	525151
DK	614	7	609	120	282	26	534	1918	1504	126	234	6573	35	145	3079	15080	39412
EE	163	0	85	25	312	1	110	327	36	34	12	586	2	15	1166	3392	11940
ES	5015	45	657	237	90	22	3754	743	2756	7162	17560	1862	66	322	29270	72850	198932
FI	291	1	40	86	54	9	374	463	156	49	38	3497	8	26	1572	3899	16142
FR	8321	36	229	6032	159	103	3500	634	6209	18530	4978	1567	174	823	49770	121600	342130
GR	995	9	35	35	17	26	1166	101	1384	160	1610	907	18	110	10360	15970	60274
HU	1586	3	14	53	20	8	1180	176	344	145	10540	857	55	1865	9035	18310	80296
IE	772	2	617	66	191	37	601	88	1477	323	536	431	12	101	18040	25780	62175
IS	69	0	56	30	28	4	91	258	526	42	13	684	10	40	212	898	4825
IT	188	49	188	635	115	246	2163	292	5484	841	14540	998	676	739	15210	39070	161449
LI	18	0	0	0	1	0	2	1	7	7	1	1	3	4	1	131	622
LT	675	0	28	868	2	2	385	1334	574	130	19	882	5	23	4351	8896	30643
LU	434	0	14	9	9	5	179	20	96	1935	76	123	22	21	405	887	17193
LV	312	1	642	16	16	3	182	347	128	30	26	445	3	20	1066	4851	13188
MT	247	0	2	7	2	2	57	12	11	20	17	51	3	14	1962	707	4171
NL	1771	2	107	290	57	56	938	2747	1201	1201	603	1458	77	307	10570	59870	125461
NO	119	1	97	21	50	6	359	391	45	45	63	5998	5	79	1195	5015	19286
PL	16240	11	354	292	165	23	9012	5788	264	186	6238	29	1009	83020	39150	334426	
PT	862	15	63	2978	28	22	1842	146	177	666	491	257	13	55	8535	13820	91509
RO	72060	3	16	177	35	14	1470	515	248	666	1587	20	1905	3062	22540	259497	
SE	604	1	119	181	120	81	753	4706	887	141	223	42	70	5281	16780	47594	
SI	675	7	4	32	4	2	134	24	27	14	11	80	71	90	5545	10125	
SK	1399	5	13	34	21	8	670	340	408	39	297	198	53	5167	7717	61347	
UK	9577	1	810	282	286	1008	6272	1640	12370	5548	918	4537	47	797	260500	478374	
RW	198300	115	5023	565	1515	850	51880	19890	20020	23900	41890	53760	19580	12200	383000	2035952	
Total	375368	865	12890	18947	6276	3170	107417	44773	174902	86227	143217	100717	24060	61629	714663	1215725	

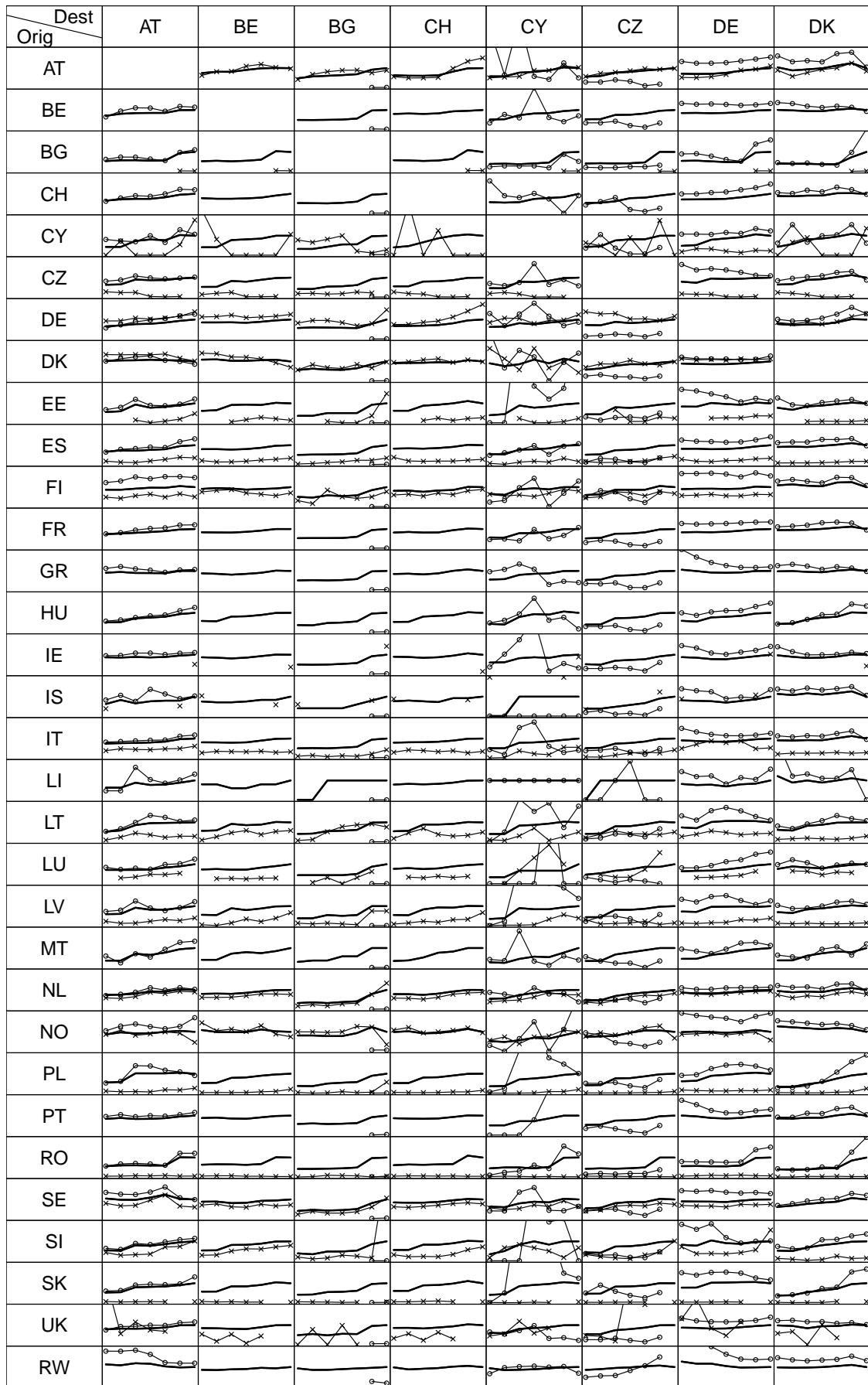


Figure A: Median estimates of the true flows (solid), reported emigration (cross) and immigration (circle) data, 2002-2008

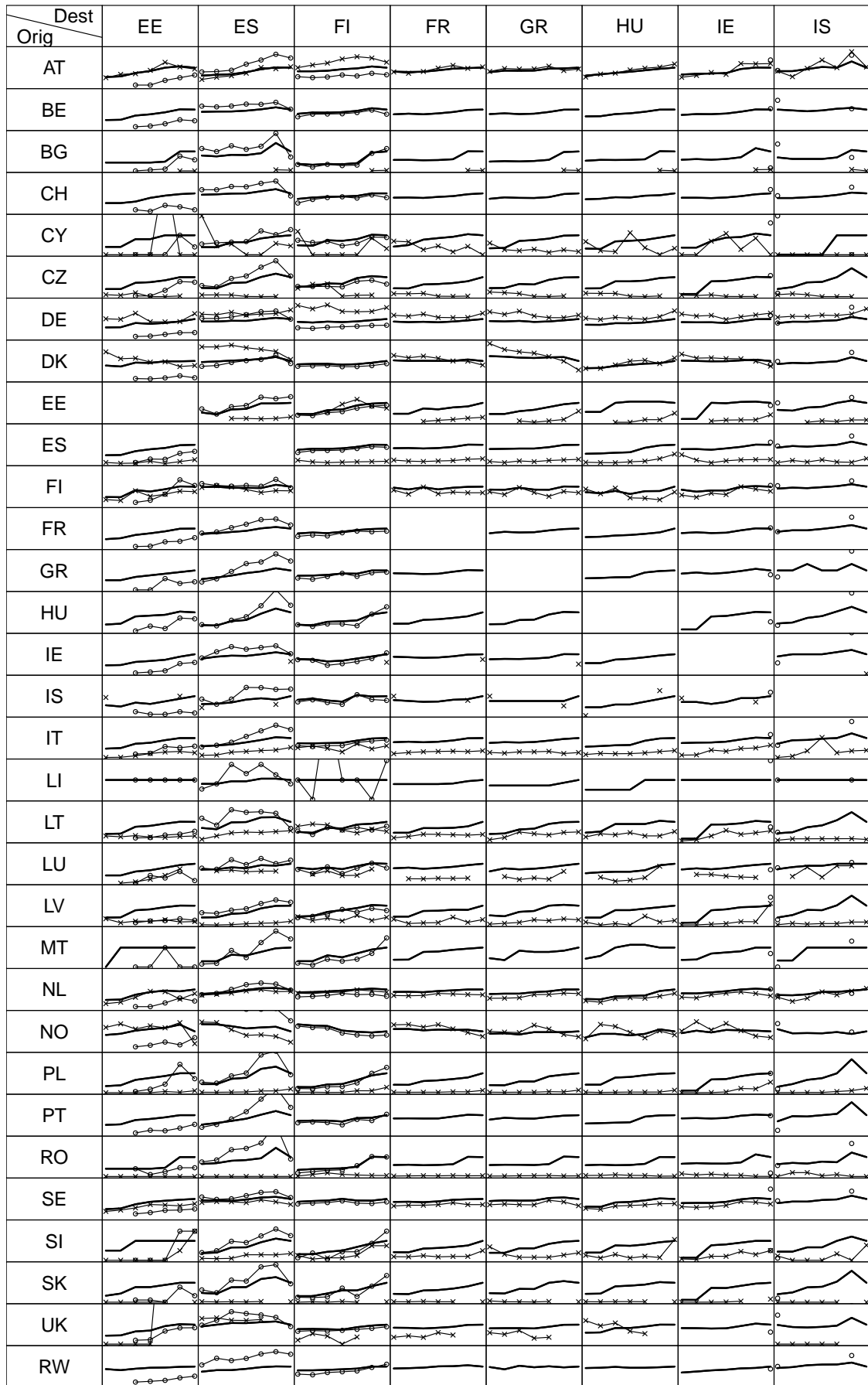


Figure A (continued): Median estimates of the true flows (solid), reported emigration (cross) and immigration (circle) data, 2002-2008

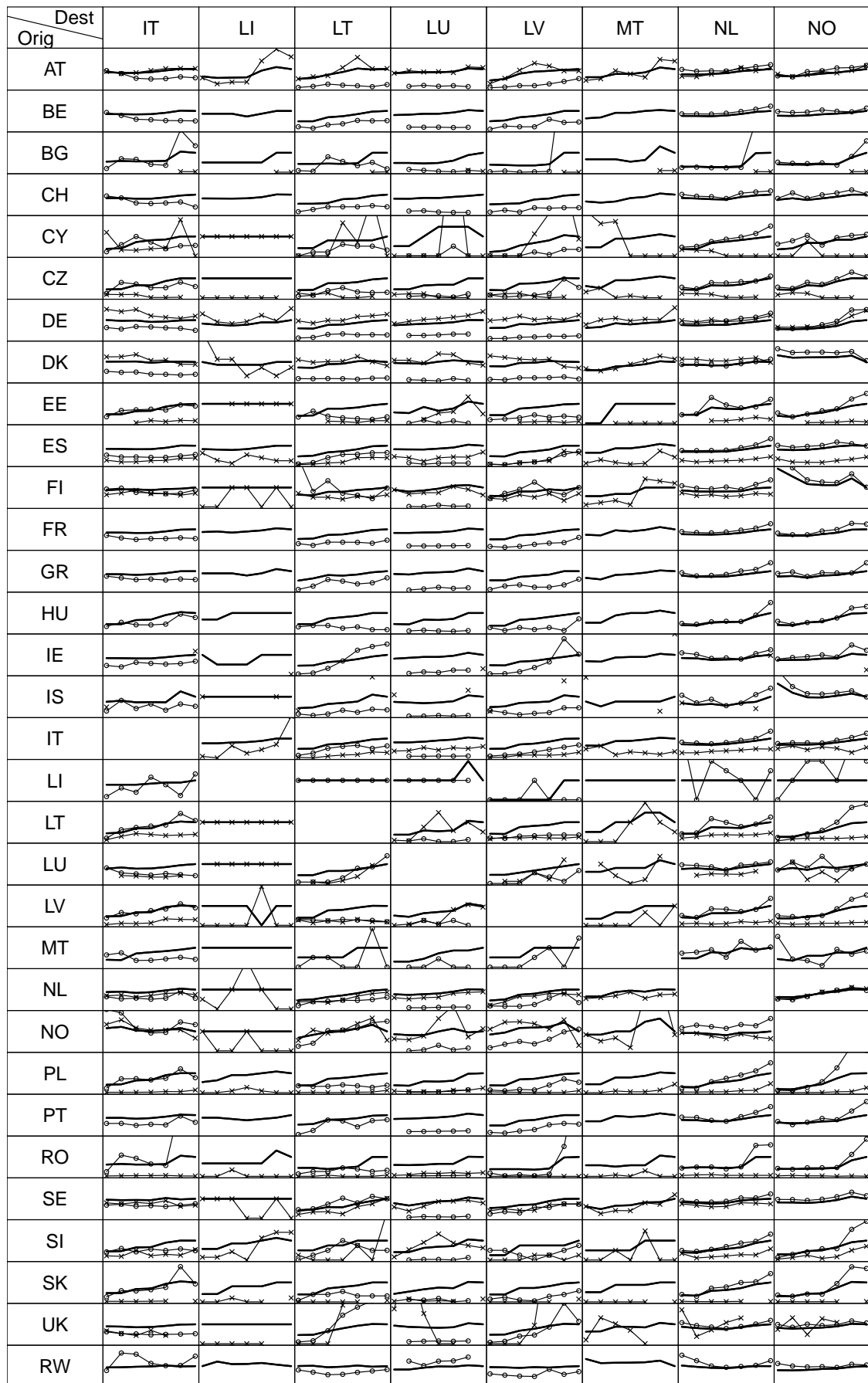


Figure A (continued): Median estimates of the true flows (solid), reported emigration (cross) and immigration (circle) data, 2002-2008

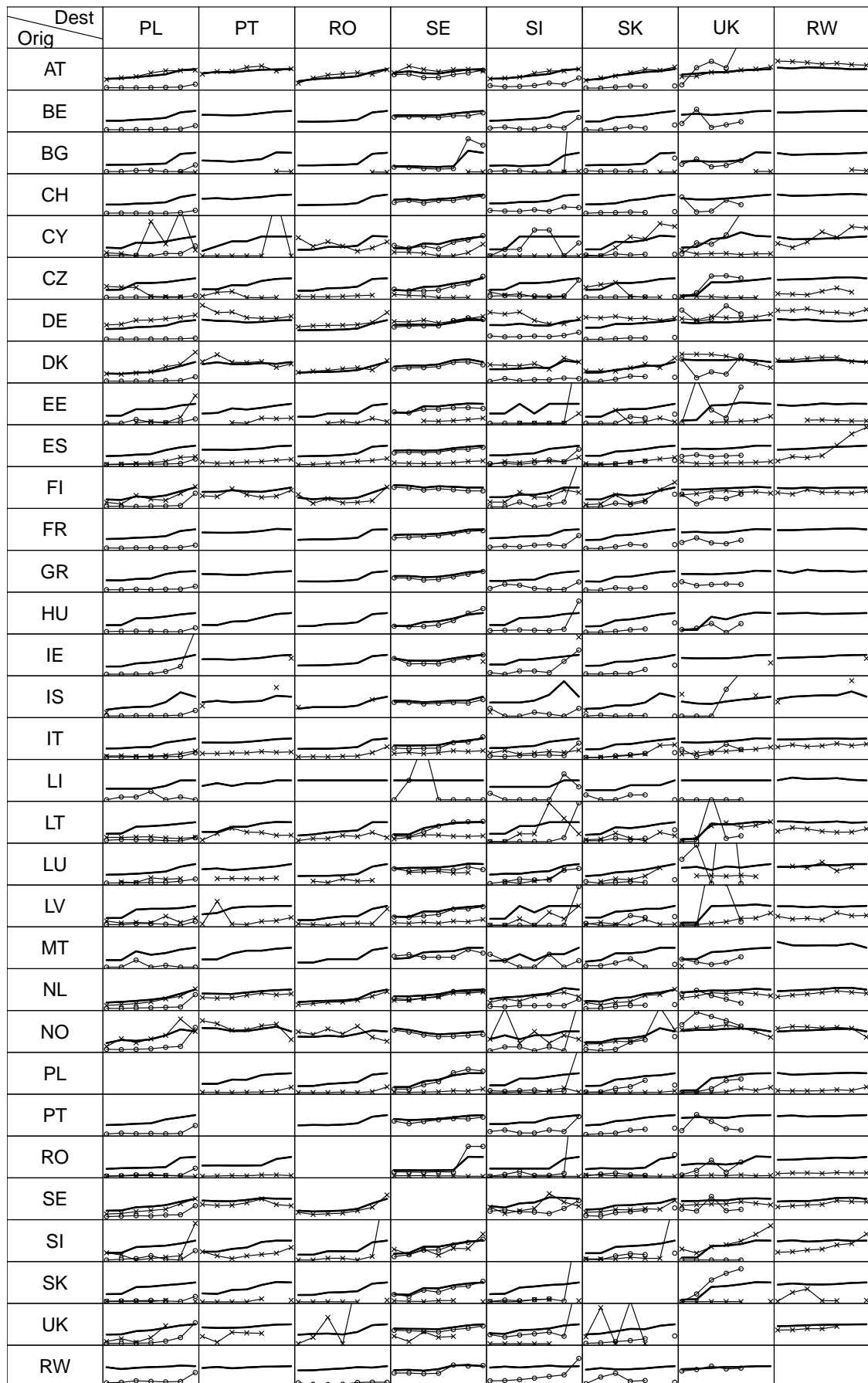


Figure A (continued): Median estimates of the true flows (solid), reported emigration (cross) and immigration (circle) data, 2002-2008