VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 583–607, http://dx.doi.org/10.2478/jos-2013-0041

Utilising Expert Opinion to Improve the Measurement of International Migration in Europe

Arkadiusz Wiśniowski[1], *Jakub Bijak*[1], *Solveig Christiansen*[2], *Jonathan J. Forster*[1], *Nico Keilman*[2], *James Raymer*[1,3], *and Peter W.F. Smith*[1]

In this article, we first discuss the need to augment reported flows of international migration in Europe with additional knowledge gained from experts on measurement, quality and coverage. Second, we present our method for eliciting this information. Third, we describe how this information is converted into prior distributions for subsequent use in a Bayesian model for estimating migration flows amongst countries in the European Union (EU) and European Free Trade Association (EFTA). The article concludes with an assessment of the importance of expert information and a discussion of lessons learned from the elicitation process.

Key words: Bayesian modelling; elicitation; expert knowledge; migration statistics; Delphi Survey.

1. Introduction

To fully understand the causes and consequences of international movements in Europe, researchers and policy makers need to overcome the limitations of the various data sources, including inconsistencies in data availability, quality and collection mechanisms. For example, in 2007, Germany reported receiving 15,515 migrants from Spain, whereas Spain only reported sending 3,601 migrants to Germany. From this single example, many questions arise: Why are the two numbers so different? How accurate are the data provided by the two countries? Could measurement be responsible for some of the difference? In this article, we describe our attempt to answer these questions by collecting information from experts on migration data.

[1] Southampton Statistical Sciences Research Institute, University of Southampton, University Road, Southampton SO17 1BJ, UK. Emails: a.wisniowski@soton.ac.uk, j.bijak@soton.ac.uk, j.raymer@soton.ac.uk, j.j.forster@soton.ac.uk and p.w.smith@soton.ac.uk
[2] Department of Economics, University of Oslo, P.O. Box 1095 Blindern, N-0317 Oslo, Norway. Emails: s.g.christiansen@econ.uio.no and n.w.keilman@econ.uio.no
[3] Australian Demographic and Social Research Institute, Australian National University, Coombs Building, Canberra AT 0200, Australia. Email: j.raymer@soton.ac.uk

Acknowledgments: This research is part of the integrated Modelling of European Migration (IMEM) project funded by the New Opportunities for Research Funding Agency Co-opertion Europe (NORFACE). We gratefully acknowledge the help of the following persons, who acted as an expert, tested the pilot survey questionnaire, or contributed otherwise to the development of the questionnaire: Guy J. Abel, Corrado Bonifazi, Harri Cruijsen, Frank Heins, Michael Jandl, John Kelly, Ewa Kępińska, Dorota Kupiszewska, Marek Kupiszewski, Giampaolo Lanzieri, João Peixoto, Nicolas Perrin, Michel Poulain, and Rob van der Erf. It should be stressed that all persons contributed in their own personal capacity to the project. We also wish to acknowledge comments received from reviewers and from the editor of this journal, which greatly improved the presentation of this article.

© Statistics Sweden

This information is gathered for use as prior inputs into a Bayesian model for harmonising and estimating international migration flows amongst the 31 countries in the European Union (EU) and the European Free Trade Association (EFTA) (Raymer et al. 2013).

Bayesian statistical methods are particularly adept at handling data from different sources and are ideal for situations in which some of the data are inadequate or missing. Additional expert information can be included in the form of prior distributions reflecting expert beliefs and judgements. The resulting estimates are then based on posterior distributions, which combine these expert beliefs with other available information, including all relevant data sources and covariates. The posterior distributions can also be used to quantify uncertainty in the estimates, providing the users, such as governments and planning agencies, with valuable additional information to design their policies directed at supplying particular social services or at influencing levels of migration (Bijak and Wiśniowski 2010).

The structure of this article is as follows. First, we describe the underlying conceptual framework for harmonising and estimating flows of international migration within Europe. Second, we outline our approach for eliciting information from experts concerning the characteristics of the reported statistics on flows. Third, we present our methodology for translating this expert information into informative prior distributions for subsequent use in the model for migration flows. We illustrate the method with an application to a European migration flow matrix for 2002–2008. The article ends with an assessment of the importance of expert information and a discussion of lessons learned from the elicitation process, followed by some conclusions.

2. A Conceptual Framework for Modelling Migration

There have been several attempts to harmonise international migration flow statistics in Europe. Poulain (1993) developed a constrained optimisation procedure to minimise the differences between two origin-destination migration flow tables representing sending and receiving country reported statistics. His 'correction factor' method has been extended more recently by Poulain and Dal (2008), Abel (2010) and De Beer et al. (2010). 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. Finally, Nowok (2010) proposed a probabilistic framework for harmonising international migration statistics (see also Nowok and Willekens 2011). Our approach to harmonising migration flows differs from these works by the emphasis on modelling the measurement aspects of the reported statistics and by providing measures of uncertainty. In this section, we introduce the underlying conceptual framework for estimating migration flows in Europe, which has been developed as a Bayesian model in (Raymer et al. 2013). In the following section, we turn to the main focus of this article: the elicitation of expert judgements.

The framework we have developed permits expert opinion to be combined with the data on migration flows and covariate information to strengthen the inference. The approach also facilitates the combination of multiple data sources, with their differing levels of error, as well as prior information about the structures of migration processes, into a single

prediction with associated measures of uncertainty. Given the substantial inconsistencies in reported statistics on international migration flows in Europe (Poulain et al. 2006), the elicitation of expert opinion concerning various aspects thereof is critical for the success of the whole modelling exercise.

In terms of measurement, true flows are assumed to be consistent with the United Nations (1998) recommendation for long-term international migration:

> 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 residence effectively becomes his or her new country of usual residence. From the perspective of the country of departure, the person will be a long-term emigrant and from that of the country of arrival, the person will be a long-term immigrant (United Nations 1998, p. 18).

Place of 'usual residence' is defined as

> The country in which a person lives, that is to say, the country in which he or she has a place to live where he or she normally spends the daily period of rest. Temporary travel abroad for purposes of recreation, holiday, visits to friends and relatives, business, medical treatment or religious pilgrimage does not change a person's country of usual residence (United Nations 1998, p. 17).

Finally, the United Nations definition we have adopted includes undocumented (irregular) migrants. In practice, the migration statistics in most countries do not cover undocumented migrants (for obvious reasons). Thus, one of the aims of the presented approach is to use expert judgement to address the levels of this aspect of migration.

Our approach to measuring migration takes into account four aspects assumed to be independent: (i) accuracy of data collection system, (ii) duration criteria used to qualify migrants that differ from the twelve months in the UN definition, (iii) undercount and (iv) coverage of migrants. Let z_{ijt}^{k} denote the counts (flows) from country i to country j during year t reported by country k, either the sending $k = i$ or receiving $k = j$. The interest of this research is to estimate y_{ijt} – the true unknown flow of migration from country i to country j in year t. It includes migration flows to and from the rest of world. Note that for each y_{ijt} there are potentially two reported flows: z_{ijt}^{i} and z_{ijt}^{j}.

We assume that the observed data z reflect the true flows y, distorted by the above mentioned deficiencies of the migration statistics, that is

$$z_{ijt}^{k} = y_{ijt} \times dur_k \times und_k \times cov_k \times err_{ijt}^{k}. \tag{1}$$

The variance of the general error term err_{ijt}^{k} measures the accuracy of the data collection system for country k. It informs the end users of the outcomes of this study on the quality of the data and measurement mechanisms utilised to collect the data. 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. Here, we distinguish three types of systems: (1) interlinked population registers in the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden), which exchange migration information; (2) other good-quality registers (The Netherlands, Germany, Austria, Belgium, Switzerland, and immigration in Spain) and

(3) less reliable registers and survey-based systems (Poland, Bulgaria, Estonia, Lithuania, Latvia, Italy, Slovenia, Slovakia, Romania, the Czech Republic, Greece, Hungary, Liechtenstein, Malta, France, Luxembourg, Portugal, United Kingdom, Cyprus, Ireland, and emigration from Spain). Our typology of accuracy is based on reports from the MIMOSA project (Kupiszewska and Wiśniowski 2009; Van der Erf 2009) and our own assessment of the data quality in Europe.

The duration parameter dur_k reflects the difference between the duration of stay criterion adopted by the country k data collection system and the baseline twelve-month criterion of the UN. For example, if a given country uses a six-month criterion, the number of true migrants (i.e., residing for twelve months or more) should be smaller than the reported number of migrants, independent of the other measurement deficiencies. Note that in practice the duration is intended or planned rather than actual.

We interpret the undercount parameter und_k as a fraction of the true flow that is captured by the data collection system in a given country. We propose two classifications here. In both of them, we work with two levels of undercount. The first one distinguishes between intra-European flows and those to and from the rest of the world. In the second one, we classify some countries as having high undercount and others as having low undercount; see Section 5 for details. The latter classification of countries with low or high undercount is based on our own assessment, as well as reports from the various projects (Poulain et al. 2006; Kupiszewska and Wiśniowski 2009; Van der Erf 2009).

The country-specific error parameters cov_k reflect the discrepancies between the observed data and the true flows that are not captured by the more general undercount parameters. These often include certain subgroups, such as international students or refugees, in the reported migration flows (Poulain et al. 2006; Kupiszewska and Wiśniowski 2009). Furthermore, we assume these parameters to lie between zero and one and interpret them as the differences in coverage with respect to the United Nations definition of migration. Given that the coverage parameters are country-specific, we assume that they measure the proportions of migration covered in relation to the true flows. For the Nordic countries and the Netherlands, these parameters are constrained to one, that is, we assume that there are no coverage errors for these countries. This assumption ensures identifiability of the parameters. For the rest of the countries, we use noninformative prior distributions. We considered the elicitation of the country-specific prior densities infeasible for the scale of our project. This approach would require at least five experts for each of the 31 countries under study. Also, since the coverage aspect of the measurement model did not utilise expert judgements, it is not discussed further in this article.

3. Obtaining Expert Information

The approach described in Section 2 requires prior information on the quality of data sources, differences in various aspects of measurement and covariates used to predict missing data. In this case, external expert judgement was sought only on the data and measurement aspects of the underlying migration flows. The experts in data collection systems were asked to rate the credibility they give to different types of migration data collected from different types of collection mechanisms (e.g., survey versus register), and to compare sending country data (i.e., emigration flows) with receiving country data

(i.e., immigration flows). Experts were also asked about the bias (e.g., systematic undercount) in the reported migration flow statistics. Each expert was asked to give us a set of values concerning certain parameters, which we then converted into probability distributions. The totality of resulting expert opinions was subsequently combined into a single set of distributions, allowing for the introduction of yet another source of uncertainty, related to the heterogeneity of experts.

To facilitate the elicitation of expert judgements, a two-stage process was used within a Delphi survey framework, whereby the expert opinions were allowed to be informed and influenced by other experts' views. This process provided a convenient avenue for the exchange of opinions and views as well as for clarifying any ambiguities as to the underlying concepts and ideas.

The elicitation of expert opinion to construct probability distributions has a long history (O'Hagan et al. 2006). In general, the acquisition of such information is a very difficult task (Kadane and Wolfson 1998). Asking an expert to draw a distribution would assume he or she has a statistical background or require us to provide such training. In our study, we could not guarantee all experts had a statistical background and did not have the time or resources to provide training. As a result and based on the feedback we received from pretesting the questionnaire, we had to limit the use of statistical terms, such as 'quantile', 'distribution', 'variance' and 'precision'. For this reason, we followed the elicitation guidelines of O'Hagan (1998) and O'Hagan et al. (2006), as well as an example of elicitation of opinion from 'non-statisticians' in Szreder and Osiewalski (1992).

From our heterogeneous group of experts, we sought basic information on particular values associated with the measurement of migration flows, which we then converted into probability distributions that could be used in our computations. After the first Delphi round, experts were provided with the densities resulting from our interpretation and

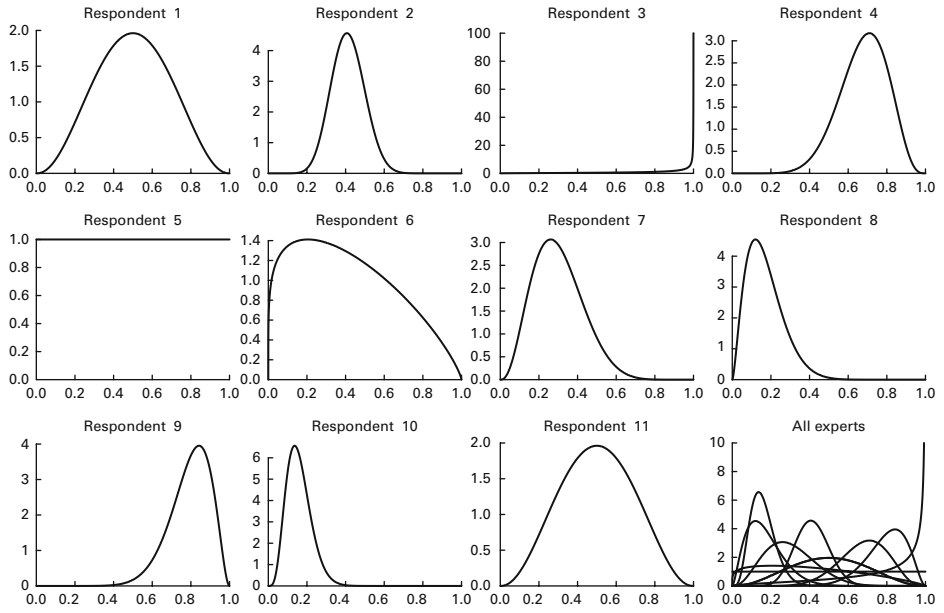

Fig. 1. Selected graphical representations of expert answers from Round 1: Undercount of emigration

parametrisation of their answers (see Figure 1 and Section 4), as well as the anonymous results from other experts in the study. This allowed them to reconsider and revise their opinions.

When formulating questions, it is important to prevent respondents from being overconfident in their opinions. For example, questions about means or medians may lead to anchoring the answer and lowering the uncertainty about the tails of the distribution (Kadane and Wolfson 1998; Rowe and Wright 2001). To avoid this problem, we constructed questions that focused on ranges of values with direct interpretations and the certainty about these ranges. Each certainty could then be interpreted as a probability that a given parameter lies within a specified range.

Experts were free to select the upper and the lower bounds of the intervals. There is an extensive literature on the issue of fixed versus variable interval bounds; see, for example, Kadane and Wolfson (1998), Garthwaite et al. (2005) or Dey and Liu (2007) for reviews. One problem with preselected intervals is that uncertainty may vary across individuals in complex ways, and hence it is difficult to find an optimal design of a preselected interval. On the other hand, lower and upper quantiles (often used in preselected intervals) have the advantage that they can be assessed by a method of bisection, as described in Garthwaite et al. (2005). From the literature on fixed and preselected intervals they also concluded that there is conflicting evidence as to which method performs better.

In one of the questions to our experts, we asked about their subjective probability concerning the accuracy of the data collection system (see Subsection 4.3). As pointed out in the literature, elicitation of probabilities is a difficult task. The perception of probability may vary depending on the formulation of the question, for example, odds ratios tend to be more extreme than the probability specified within a range [0, 1] (Goodwin and Wright 1998). Another issue is viewing uncertainty in terms of frequencies rather than subjective probabilities (Gigerenzer 1994; Kadane and Wolfson 1998) and forgetting about the context of an event under consideration. Hence, in the formulation of our question, we followed the advice of Gigerenzer (1994) of asking about proportions and providing the context of the subject.

3.1. Delphi Technique

The Delphi technique is a method used to obtain information from a group of experts in order to make judgements and forecasts when extensive or reliable data in the field of enquiry are not available (Rowe and Wright 1999). It was first developed by the RAND Corporation for US military use in the 1950s. More recently, and in the context of international migration in Europe, this technique was applied to (i) forecast migration between Central and Western Europe after the fall of communism (Drbohlav 1996), (ii) the MIGIWE (Migration and Irregular Work in Europe) project to gain information on irregular foreign employment in Austria following the 5th Enlargement of the EU (Jandl et al. 2007) and (iii) the IDEA (Mediterranean and Eastern European Countries as new immigration destinations in the European Union) project to augment forecasting models for seven European countries (Wiśniowski and Bijak 2009; Bijak and Wiśniowski 2010).

In a Delphi survey, the elicitation of expert opinions takes the form of an anonymous questionnaire with multiple rounds, where the experts report their subjective beliefs on the

topics in question. Between rounds, experts are provided with feedback on the answers in the preceding round, including qualitative arguments in support of various views. The experts then complete the next round of the survey where they are free to alter their previous answers in light of the new information provided by the feedback.

According to Rowe and Wright (2001), the Delphi technique is most reliable when there are between five and 20 respondents who are experts in the field of enquiry and when there is heterogeneity among the experts. The questions should be sufficiently comprehensive to contain the relevant information but not cause information overload. The final round answers are usually weighted equally. Past evaluations have shown that the answers from the final round Delphi surveys are more accurate than other approaches using only one expert, focus groups or single-round questionnaires. By using an anonymous questionnaire instead of a group meeting, one avoids group pressure and the domination of the group by some individuals. The Delphi method may also lead to better results because the experts think more carefully when responding when they know that their answers will be given as feedback to other experts.

3.2. Constructing the Questionnaire

For our project, the elicitation process consisted of two rounds (hereafter Round 1 and Round 2) and involved eleven external experts. We selected the experts from among those international colleagues who we thought would be knowledgeable about the measurement of international migration in several countries. The online questionnaire was pretested by an additional two external experts and two of our team members. The survey was preceded by an invitation letter, in which the aim of the project and the purpose of the questionnaire were explained. The experts were asked to give their opinion about how specific measurements of international migration deviate from the benchmark of the United Nations definition of a long-term migrant (see Section 2).

The Round 1 questionnaire included a definition of a long-term migrant according to the United Nations definition discussed above plus 14 questions grouped into four sections. Each section contained a specific set of closed questions and an open question, in which experts were allowed to express their comments or arguments related to their answers. In all questions, experts were asked to provide their answers in terms of percentages, and to state how certain they were about their answers, that is, 50%, 75%, 90%, 95% or Other. The first three sections of the questionnaire were restricted to intra-EU/EFTA migrants, while the fourth section concerned migration between the EU/EFTA countries and the rest of the world. Finally, the experts were also allowed to provide general comments or suggestions, as well as to ask questions of their own. The full questionnaire is available for download at [http://www.imem.cpc.ac.uk].

The undercount of migration between EU and EFTA countries and from or to the rest of the world was the focus of Section A (Questions 1–3) and Section D (Questions 12–14) of the questionnaire respectively. Here, experts were asked to provide their judgements and uncertainty regarding the lowest and highest percentages of the possible undercount of emigration and immigration in the published statistics. To do this, the experts needed to consider a nonspecific, hypothetical European country with a good population register and migration definitions corresponding exactly with the United Nations (1998)

recommendation. In other words, the experts were asked to think of migration collection systems rather than specific country experiences.

The focus of Section B of the questionnaire (Questions 4–6) concerned the duration of stay criteria included in the definition of migration. In Europe, different timing criteria are used by different countries and these questions aimed at assessing how this might affect the relative levels of reported migration. Thus, in Question 4, experts were asked how much, in percentage terms, the level of migration would be for a duration of stay criterion of six months instead of twelve months. Question 5 asked for the difference between three- and six-month criteria.

Finally, the questions in Section C were aimed at obtaining information about the accuracy of population registers in measuring migration. Experts were asked to consider registers in which there was no systematic bias and with random factors being the main source of error. In Questions 7 to 11, experts were asked to provide their beliefs and certainty regarding published statistics falling within an interval from minus 5% to plus 5% compared to the true total level of emigration and immigration.

All eleven respondents from Round 1 took part in Round 2 of the survey. Of these, nine chose to change their answers to one or more of questions in Round 2. Further information about the changes in the experts' opinions between the two rounds can be found in the following section. The questionnaire in Round 2 consisted of the same set of questions as in Round 1. It also contained anonymised answers from Round 1 and the arguments used to support the various views, including the underlying reasons for different assessments. The experts also had the option to look at graphical representations of their individual answers, examples of which are shown in Figure 1. Details on how these distributions were compiled are provided in Subsection 4.1.

4. Translating the Expert Information into Prior Distributions

In this section, we explain how the opinions and judgements obtained in the first and second round of the Delphi survey were translated into prior distributions for the parameters introduced in Section 2. The parameters in question are used to address undercount, duration of stay and accuracy of measured migration flows.

The construction of prior densities based on expert answers was a three-step process. First, having obtained the raw answers to a given question about some parameter θ, we identified a distribution, that, in our opinion, reflected the expert judgements about the θ most appropriately. Second, we constructed a prior density $f_i(\theta)$ for each expert i, $i = 1, \ldots, n$. Third, we combined the individual densities into a single prior density:

$$P(\theta) \sim \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \tag{2}$$

We chose to have an equally-weighted opinion pool because it allowed us to have a simple, robust and general method for aggregating expert knowledge. Aggregation methods based on weighting, such as that of Cooke (1991), require a separate elicitation round in which each expert is asked about a particular variable, of which the real value is known to the facilitator but not to the expert. In our situation, we did not know the real values of any of the parameters. Therefore, we assigned equal weights to the experts. The

equal weights also allowed the different and sometimes opposing assessments to be fed into the estimation model. Smoothing techniques or fitting a parametric distribution to the expert answers, for example, would have reduced the amount of information provided by the experts. Another option, which could be explored in the future work, would be to perform Bayesian model averaging over models with each single expert prior distribution as a separate input. For a discussion about the benefits and consequences of the various ways expert opinions can be combined, we refer the reader to Clemen and Winkler (1990) and O'Hagan et al. (2006).

4.1. Undercount of Emigration and Immigration

4.1.1. Method for Constructing the Prior Density

In the first and fourth section of the Delphi questionnaire, experts were asked to provide answers to the following question about undercount of migration within Europe and to and from the rest of world. In the preamble to the question on undercount, the reference to the baseline UN definition was made. The question was formulated as:

> [. . .] Consider a European country with a good population register, e.g., Sweden or Finland, that has fully adopted the UN definition. Because migrants do not always have sufficient incentives to report their moves to the relevant authorities, migration statistics are often lower than the true total level. For immigrants this difference is thought to be smaller than for emigrants.
>
> (a) *By how many per cent do you expect that emigration (or immigration) flows are undercounted in the published statistics, as compared to the true total level of emigration (immigration)? Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true undercount will lie within the range that you provided above?*

Let P_1 and P_2 denote the lower and upper percentages stated by an expert about undercount and c denote the certainty about the range (P_1, P_2). The underlying assumption regarding undercount is that $P \in [0, 1] \times 100\%$, which is

$$(1 - P)y = z, \tag{3}$$

where y are true flows and z are reported flows. Then $(1 - P)$ can be interpreted as a fraction of the true flow which is captured in the reported data. A couple of the answers provided by experts in the first round were not meaningful, suggesting some difficulties were experienced in interpreting the questions. We addressed this issue in the Round 2 questionnaire (see the following section).

To convert the experts' answers into prior distributions for the parameters, we first had to identify which probability distributions would both accurately reflect experts' beliefs and work well with the underlying conceptual framework introduced in Section 2. We considered three densities: piecewise uniform, logit-normal and beta. These densities were chosen because they could be constrained to values between zero and one and they were flexible in terms of shapes. Besides, as opposed to truncated distributions such as normal or log-normal, their parameters could be easily calculated.

Table 1. Experts answers to question 1 – undercount of emigration

Respondent	1	2	3	4
Lowest percentage, P_1	20	30	50	4
Highest percentage, P_2	80	50	90	8
Certainty, c	90	75	90	5

To illustrate the differences between various densities, consider four answers of the experts to Question 1 set out in Table 1. For example, Respondent 2 believes that the emigration flows in the published statistics are undercounted by 30% to 50% with a probability of 75%. Respondent 4, on the other hand, believes that the reported flows of emigrants are only 4% to 8% too low, which represents a very precise range, but his or her certainty is only 5%. It should be intuitive that the wider the range of undercount, the larger the certainty should be. Note that in Round 1 of the Delphi survey, almost all answers were consistent with this rule. For the questions concerning undercount, only one expert indicated relatively large range with a small level of certainty. This led to some computational and interpretation problems.

For the case of the piecewise uniform densities, the computation was straightforward. We assumed that the certainty level c provided by a given respondent corresponded with the probability mass between P_1 and P_2. The remainder, $(1 - c)$, was proportionally distributed between $[0, P_1]$ and $[P_1, 1]$. Thus the quantiles of the resulting piecewise uniform density were

$$q_1 = \frac{(1-c)P_1}{1 + P_1 - P_2} \quad \text{and} \quad q_2 = \frac{(1-c)(1-P_2)}{1 + P_1 - P_2}. \tag{4}$$

The resulting piecewise uniform densities, after transformation into undercount using Equation (3), are presented in the first row of Figure 2.

In the case of the logit-normal density, it was assumed that

$$\begin{cases} \mu + \sigma\,\Phi^{-1}(q_1) = \dfrac{\log(P_1)}{1 - \log(P_1)} \\ \mu + \sigma\,\Phi^{-1}(q_2) = \dfrac{\log(P_2)}{1 - \log(P_2)} \end{cases} \tag{5}$$

where μ and σ are expected value and standard deviation of the underlying normal density and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution. Two specifications of q_1 were considered. In the first one, the probability mass c lies between P_1 and P_2 and the remainder, $(1 - c)$, symmetrically distributed between $[0, P_1]$ and $[P_2, 1]$:

$$q_1 = \frac{1-c}{2} \quad \text{and} \quad q_2 = \frac{1+c}{2} \tag{6}$$

The second specification is based on quantiles as in the piecewise uniform approach, as given by Equation (4). The resulting densities for these two approaches, after

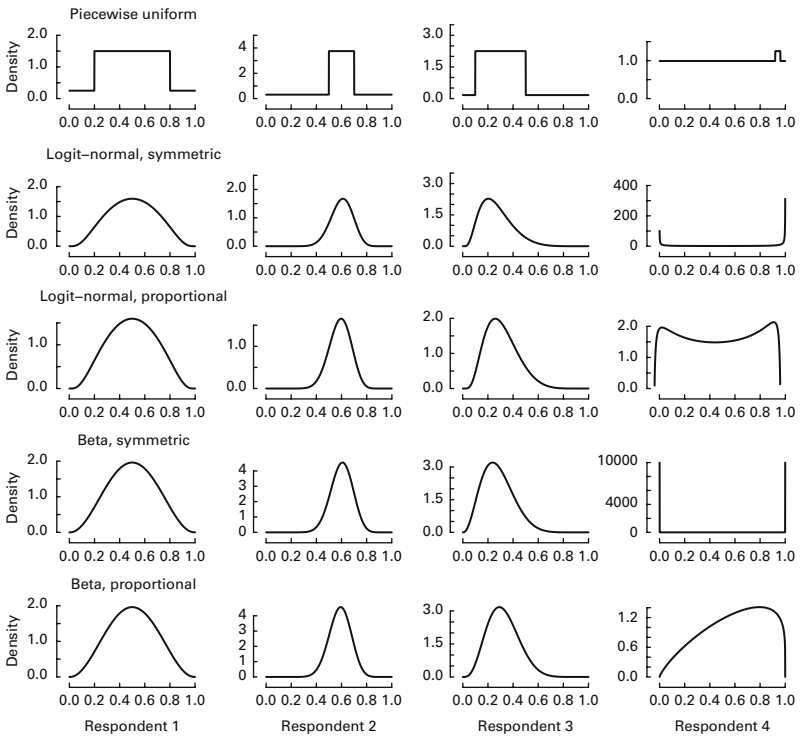

Fig. 2. Densities for four experts with various specifications

transformation using Equation (3), are presented in second and third row of Figure 2 respectively.

Finally, two sets of quantiles were also considered for the beta distribution. The parameters α and β of the beta density were computed by solving a set of two equations:

$$\begin{cases} F_b^{-1}(P_1, \alpha, \beta) = q_1 \\ F_b^{-1}(P_2, \alpha, \beta) = q_2 \end{cases}, \tag{7}$$

where F_b^{-1} is an inverse cumulative distribution function of the beta distribution. This was achieved by finding roots of the following expression:

$$\sum_{i=1}^{2} \left[F_b^{-1}(P_i, \alpha, \beta) - q_i\right]^2, \tag{8}$$

where q_1 and q_2 were either proportionally (4) or symmetrically (6) distributed. Vector $(\alpha_0 = 1, \beta_0 = 1)$ was used as a starting point for this algorithm. The densities obtained for the four example experts are presented in Figure 2 in the fourth and fifth rows for symmetric and proportional quantiles respectively.

From all of the approaches considered to translate and represent the subjective expert opinions, the beta density with proportional quantiles was ultimately chosen. Piecewise uniform was rejected because it produced relatively crude results. The logit-normal and

beta distributions with symmetric quantiles also tended to yield unintuitive shapes, especially in cases where experts assigned more certainty to regions close to zero or 100% undercount. Such a case is represented by Respondent 4 in Figure 2. Both symmetric approaches (logit-normal and beta in rows 2 and 4, respectively) are bimodal with most of the probability mass assigned close to zero and one, which was considered to be a rather implausible representation of an expert's opinion. The proportional logit-normal approach also resulted in a bimodal density and was rejected (depending on relative sizes of μ and σ, the logit-normal distribution has one or two modes; see Johnson 1949, pp. 158–159).

4.1.2. Feedback to Experts and Round 2 Questionnaire

As mentioned in Subsection 3.2, the second round of the Delphi survey included anonymised answers from the first round, together with arguments used to support the views and reasoning of various experts. Besides this feedback, we also took advantage of Round 2 to ensure a shared understanding of all underlying concepts among the participants. For example, in Round 1, a few of the experts gave answers to some of the questions on undercount which lay outside the 0–100% range, making interpretation difficult in terms of Equation (3). This suggests that the undercount was understood as 'how many times larger are the true flows, in comparison to the reported data', that is,

$$y = (1 + \alpha)z \tag{9}$$

where y and z are the true flows and reported data, respectively, and α denotes magnitude of how many times the true flows are larger than the reported data. Hence, if an expert provided at least one number α falling outside of a range [0, 1], both answers were treated according to the interpretation implied in Equation (9) and recomputed to be $P = 1 - 1/(1 + \alpha)$, where P is the undercount factor as in Equation (3). Those experts who in Round 1 had provided answers outside the 0–100% range were contacted to confirm that our interpretation of their answers was correct. In Round 2, it was specifically stressed for some of the questions that the answer must lie in the interval 0–100%.

4.1.3. Expert Answers and Resulting Prior Densities

The answers provided by the experts to the question on undercount of emigrants within EU and EFTA countries, converted into proportions, are presented in Table 2. For the

Table 2. Experts' answers concerning undercount of emigrants

Resp.	1	2	3	4	5	6	7	8	9	10	11
Round 1											
P_1	0.20	0.30	0.00	0.50	0.10	0.04	0.10	0.01	0.80	0.05	0.20
P_2	0.80	0.50	10.00	0.90	0.30	0.08	0.40	0.30	0.95	0.20	0.80
c	0.90	0.75	0.50	0.90	0.20	0.05	0.75	0.90	0.50	0.75	0.90
Round 2											
P_1	0.25	0.30	0.10	0.50	0.10	0.04	0.20	0.01	0.50	0.50	0.30
P_2	0.75	0.50	1.00	0.70	0.30	0.08	0.50	0.50	0.75	0.90	0.90
c	0.90	0.75	0.50	0.75	0.50	0.05	0.50	0.90	0.75	0.90	0.90

Resp. – Respondent, P_1 – Lowest proportion, P_2 – Highest proportion, c – Certainty.

emigration undercount we observe that two respondents did not change their opinions between two rounds of the study, while three increased their confidence. Some of the experts provided wide percentage spans with large confidence (e.g., Respondents 1, 4, 10, 11), while others gave a comparatively narrow range with lower certainty (Respondents 2, 6 and 9). Respondent 3 provided a percentage range exceeding the envisaged 0–100% range with a relatively small confidence. Hence, we interpreted it as the undercount given in Equation (9) and transformed it accordingly. In the Round 2 answers, we observe that only two experts lowered their certainty.

In Figure 3 and Figure 4, we present the Round 1 and Round 2 expert opinions regarding factors $(1 - P)$, that is, the parameters und_k which capture the emigration and immigration undercount, respectively, transformed into beta densities with proportional quantiles. The individual curves were used to construct mixed prior densities (bold curves in Figure 3 and Figure 4) for the und_k parameters.

The prior density for emigration undercount, based on answers from Round 1 (bold curve in the left plot of Figure 3), is weakly informative in the sense that there is no clear region of undercount that would be indicated by the majority of experts. The resulting density has four modes. Mean undercount is 52%, with a standard deviation of 27%. The corresponding Round 2 prior density is unimodal, with a mean of 56% and a standard deviation of 22%. Unimodality and lower spread in the second round suggests there has been some convergence of the answers.

Comparing the prior densities of the immigration undercount answers with those of emigration, we observe a shift of the probability mass from the region of a very high undercount (near zero) to the values suggested by the majority of experts, that is around 60–80%. The Round 1 prior density mean is 68% with standard deviation of 25%; in the second round these values changed to 72% and 18%. Again, the three modes of the Round 1 prior were replaced by a unimodal density in Round 2, which is a sign of convergence in judgements.

The overall large standard deviation and a relatively 'flat' shape of the distribution of the mixture densities reflects the heterogeneity of expert judgements about the undercount. It may also stem from different experiences of the experts with migration statistics. That is, their opinions may have been based on the systems known best to them or on their lack of knowledge regarding other systems.

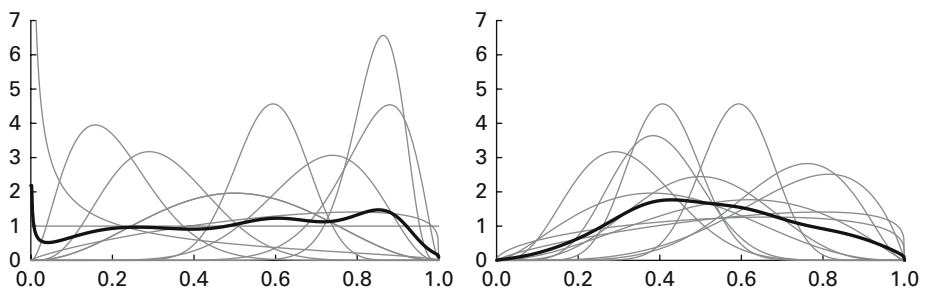

Fig. 3. Expert answers transfomed to densities for undercount of emigrants parameter, Round 1 (left) and Round 2 (right)

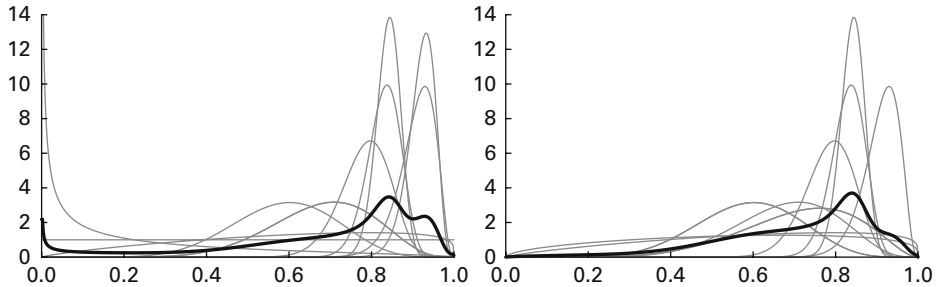

Fig. 4. Expert answers transfomed to densities for undercount of immigrants parameter, Round 1 (left) and Round 2 (right)

As shown in Figure 5 and Figure 6, the expert assessments of the undercount of emigration to and immigration from the rest of the world are more ambiguous than for intra-European migration. Four experts stood by their Round 1 answers in Round 2 and two reduced their confidence and changed the undercount range.

Consensus among experts concerning the undercount of rest of world flows was not reached. Respondents pointed out that the data on non-EU citizens are in general better captured due to more requirements for them than the data on nationals or other EU citizens. This would reduce the undercount. On the other hand, including the undocumented migrants in our estimates has had a reverse effect and blurs its evaluation.

4.2. Overcount Due to Different Duration of Stay Criteria

4.2.1. Method for Constructing the Prior Density

The duration of stay parameters capture the effects of different timing definitions used to qualify migrants. We assume that, in the presence of no undercount and the same accuracy, the shorter the duration measure, the greater the number of migrants:

$$y_p < y_{12} < y_6 < y_3 < y_0, \tag{10}$$

where the subscripts of the true flow y denote the durations with p = permanent, 12 = twelve months, 6 = six months, 3 = three months and 0 = no time limit. For

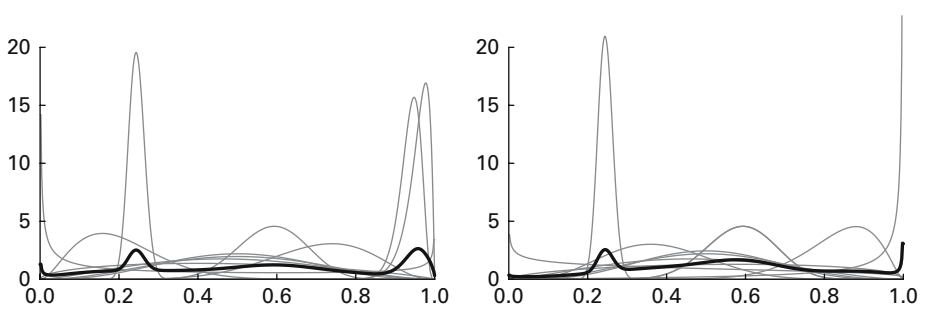

Fig. 5. Expert answers transformed to densities for undercount of emigrants to rest of world parameter, Round 1 (left) and Round 2 (right)

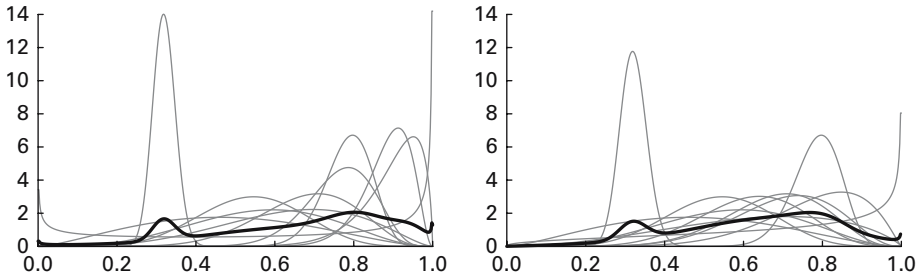

Fig. 6. Expert answers transformed to densities for undercount of immigrants to rest of world parameter, Round 1 (left) and Round 2 (right)

simplicity, we suppress country and time subscripts. Our benchmark criterion was twelve months, following the United Nations (1998) definition described in Subsection 3.2. The overcount of the number of migrants, due to the different duration criterion in the reported data z, can be expressed by a factor dur_s in the equation

$$z = dur_s \times y_{12},$$

where s denotes the applied duration criterion, that is $s \in \{0, 3, 6, 12, p\}$.

The question in the Delphi study about the overcount was introduced after the question concerning the undercount. In the preamble it was pointed out that the undercount did *not play a role in here*. It was formulated as follows:

> [. . .] Consider a European country that uses a 12-month criterion. Now imagine that the six-month criterion is used instead. With this new criterion, more persons are considered migrants compared to the previous criterion.
>
> (a) *By how many per cent do you expect that the level of migration with the SIX (THREE) MONTH criterion is higher than with the twelve (SIX) MONTH criterion? Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true value will lie within the range that you provided above?*

The experts were asked to provide lower and upper percentages of the overcount, denoted by P_1 and P_2, as well as their level of certainty about the range (P_1, P_2). The percentage $P > 0$ provided by experts represents the duration overcount in the following way:

$$y_\alpha = (1 + P)y_b, \tag{11}$$

where α denotes a shorter duration criterion than b. The overcount due to using a six-month criterion instead of a twelve-month criterion is captured by $1 + P = \exp(d_3)$, where $d_3 > 0$ is an auxiliary variable, so that $y_6 = \exp(d_3)y_{12}$. Similarly, the overcount of migrants measured using a three-month criterion compared to a six-month criterion is $\exp(d_2)$, $d_2 > 0$, which can be expressed as $y_3 = \exp(d_2)y_6$. Thus the effect of using a

three-month criterion compared to a twelve-month criterion is $y_3 = \exp(d_2 + d_3)y_{12}$. For permanent duration the relevant scaling factor is $y_P = \exp(-d_4)y_{12}$, where $d_4 > 0$. These formulations led to the following constraints imposed on the duration parameters dur_s, $s \in \{0, 3, 6, p\}$:

$$\begin{aligned} dur_0 &= \exp(d_1 + d_2 + d_3), \\ dur_3 &= \exp(d_2 + d_3), \\ dur_6 &= \exp(d_3), \\ dur_p &= \exp(-d_4). \end{aligned} \tag{12}$$

We further assume that each d_l, $l = 1, 2, 3, 4$, follows a log-normal distribution. Then the parameters of each expert-specific density for d_l can be calculated by solving the following set of equations:

$$\begin{cases} \mu + \sigma\,\Phi^{-1}(1/2 + c/2) = \log\log(1 + P_1) \\ \mu - \sigma\,\Phi^{-1}(1/2 + c/2) = \log\log(1 + P_2) \end{cases}, \tag{13}$$

where μ and σ are the expected value and standard deviation respectively of the underlying normal density, c is the elicited certainty level, and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution.

The comparisons of the 'permanent' and twelve-month criterion, as well as the three months with 'no time limit', were elicited from the migration experts during a workshop organised by the authors. This workshop brought together academics and persons responsible for migration data at national and international institutions, including some of the experts from the Delphi study. For elicitation, the same approach and formulation of the questions were used but the number of experts was 24 instead of eleven. Here we present the results only of the original Delphi questionnaire, as it is consistent with the other questions on undercount and accuracy.

4.2.2. Expert Answers and Resulting Prior Densities

The representations of individual expert answers concerning the overcount of migration due to different duration of stay criteria are presented in Figure 7 and Figure 8 for six months versus twelve months and three months versus six months respectively on the

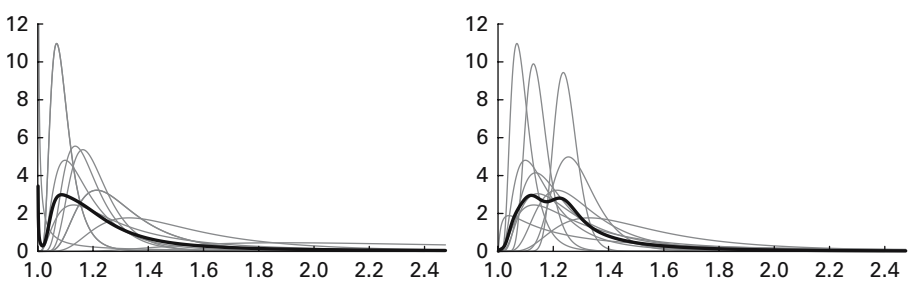

Fig. 7. Expert answers transformed to densities for duration overcount exp(d_3), 6 months versus 12 months, Round 1 (left) and Round 2 (right)

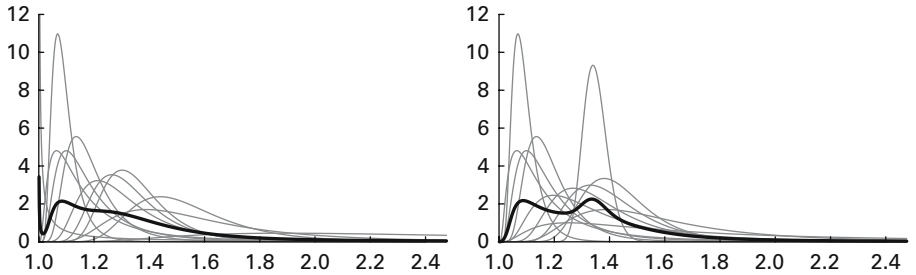

Fig. 8. Expert answers transformed to densities for duration overcount exp(d_2), 3 months versus 6 months, Round 1 (left) and Round 2 (right)

linear scale. In other words, the curves represent the expert answers translated into densities for parameters exp(d_l) and not the overcount factors dur_s.

When we compare the mixture prior densities (bold curves in Figure 7 and Figure 8) resulting from two rounds of questions about the overcount due to different duration criteria, we observe two important changes between Round 1 and Round 2 of the Delphi survey. In both the twelve month to six month and six month to three month comparisons, the expert whose answer contributed to the mode at 0% changed his or her judgement. The mixture is a heavy-tailed distribution because Respondent 3 provided a comparatively small confidence in the answers. Here, the number of migrants captured by the data collection system with six months duration of stay criterion is expected to be 10–30% larger than with the twelve-month criterion. Experts were more uncertain and ambiguous about the difference between the three- and six-month criteria.

4.3. Accuracy

4.3.1. Method for Constructing the Prior Density

The question regarding accuracy of data collection appeared to be the most challenging for the experts to answer. It was asked for in the third section of the Delphi questionnaire. In the preamble to the question, it was explained that accuracy should be assessed assuming there were no biases in the measurement, that is, it was independent from the undercount and duration issues.

> [. . .] Consider a European country with a population register in which there is no systematic bias in the measurement of migration. In this case, we may expect random factors, for instance administrative errors in the processing of the data, to affect the level of migration that is actually measured.
>
> (a) *For EMIGRATION (IMMIGRATION), how probable do you think it is that the published statistics are within an interval from minus 5% to plus 5% compared to the true total level of emigration? (If it helps, think of how often the annual published statistics are within this interval during a period of 100 years). Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true value will lie within the range that you provided above?*

The interpretation of the question in brackets was provided to help respondents understand the notion of accuracy and provide a context of the range of minus 5% to plus 5%.

To transform experts' answers into prior densities for the precision of the random terms in the measurement equations, consider a simplified equation for the observed data z and true flows y:

$$z = y \times \xi, \tag{14}$$

where ξ denotes an error term. On the logarithmic scale, ξ is normally distributed with mean zero and precision τ. Given the $\pm 5\%$ deviation from the true level of migration and two probabilities of such an event provided by the experts, P_1 and P_2, it follows that

$$P_i = \Phi[\log(1.05)\sqrt{\tau_i}] - \Phi[\log(0.95)\sqrt{\tau_i}], \quad i = 1, 2. \tag{15}$$

Using the approximation $\log(1.05) \approx -\log(0.95) \approx 0.05$, we simplify the above equation into

$$P_i = 2\Phi(0.05\sqrt{\tau_i}) - 1, \quad i = 1, 2. \tag{16}$$

Then the precision τ_i is computed as

$$\tau_i = 400\left[\Phi^{-1}\left(\frac{P_i + 1}{2}\right)\right]^2, \; i = 1, 2. \tag{17}$$

For expert-specific distribution of τ_i a gamma $\mathcal{G}(\alpha, r)$ density is assumed. Parametrisation of the gamma distribution throughout this article is such that the expected value is α/r and the variance is α/r^2. We can estimate the parameters α and r by solving the following set of equations:

$$\begin{cases} F_g^{-1}(P_1, \alpha, r) = q_1 \\ F_g^{-1}(P_2, \alpha, r) = q_2 \end{cases}, \tag{18}$$

where F_g^{-1} is an inverse cumulative distribution of the gamma distribution. This is achieved by finding the roots of the expression:

$$\sum_{i=1}^{2}\left[F_g^{-1}(P_i, \alpha, r) - q_i\right]^2, \tag{19}$$

where

$$q_1 = \frac{(1-c)P_1}{1 + P_1 - P_2} \quad \text{and} \quad q_2 = \frac{(1-c)(1-P_2)}{1 + P_1 - P_2}$$

For the cases where experts provided zero or 100% probabilities, this formula cannot be used because it has no unique solution. To overcome such answers, we replaced zeros with 0.01% and 100% with 99.99%.

To find starting point values for the optimising algorithm a log-normal approximation was used, with parameters μ and σ calculated as

$$\sigma = \frac{\log(\tau_2) - \log(\tau_1)}{\Phi^{-1}(1 - q_2) - \Phi^{-1}(q_1)} \tag{20}$$

and

$$\mu = \log(\tau_2) - \sigma\Phi^{-1}(1 - q_2). \tag{21}$$

Then, the expected value and the variance of the approximating log-normal density were computed as follows:

$$E(\tau) = \exp(\mu + \sigma^2/2)$$
$$\mathrm{Var}(\tau) = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$$

Finally, we solved the basic equations $E(\tau) = \alpha/r$ and $\mathrm{Var}(\tau) = \alpha/r^2$ for α and r to obtain the starting point values.

4.3.2. Expert Answers and Resulting Prior Densities

In Figures 9 and 10, the graphical representations of expert answers for emigration and immigration respectively are shown. For clarity, we present the densities for the expected proportion of observations with less than 5% error, as was requested in the question, rather than the gamma densities for the precision τ. The bold curves represent mixtures of the experts' single densities. In terms of results, we observe that in both Round 1 and Round 2, the experts' answers were diversified. About a third of all experts provided low probabilities suggesting that the measurement of both emigration and immigration is rather poor, while the rest of experts stated that the data collection systems are mostly accurate with probabilities higher than 50%. This heterogeneity could stem from the different backgrounds and experiences with various data collection systems in Europe.

Although experts perceived the measurement of immigration to be more accurate than emigration, their opinions were far from unanimous. For example, one of the experts, having seen the results of Round 1, reduced his or her level of confidence in Round 2. In general, we observed some convergence in opinion for the accuracy of immigration.

5. Importance of Expert Information

As described in Subsection 4.1.3, the elicited prior densities for undercount were varied and uncertain. In our process of assessment, we came to the conclusion that our original specification for the undercount parameters had likely created some confusion amongst the experts related to the difficulty in distinguishing undercount amongst intra-European flows

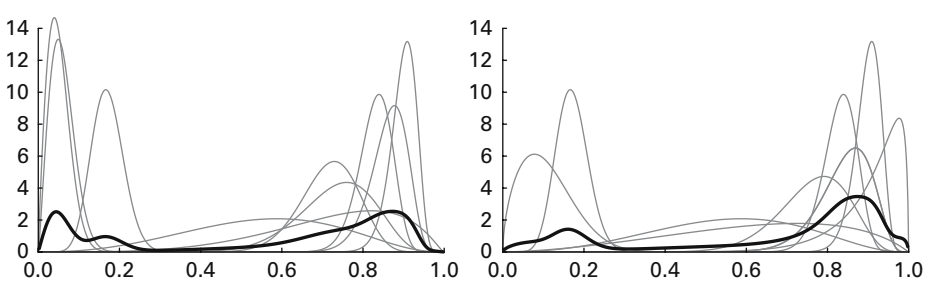

Fig. 9. Expert answers transformed to densities for accuracy of emigration measurement, Round 1 (left) and Round 2 (right)

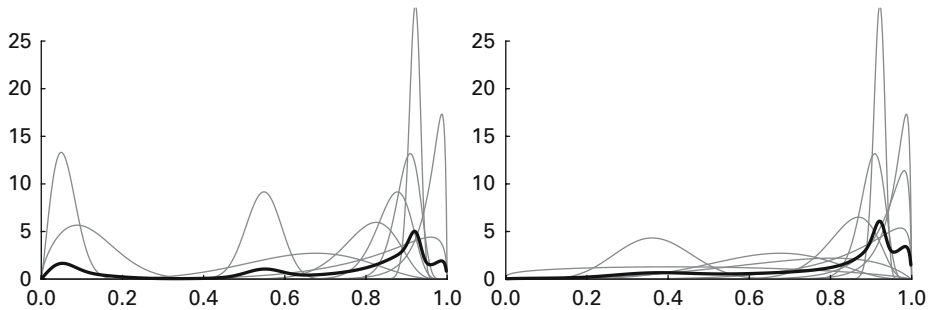

Fig. 10. Expert answers transformed to densities for accuracy of immigration measurement, Round 1 (left) and Round 2 (right)

and flows to and from rest of the world. Moreover, by running the model in (Raymer et al. 2013), we found that the prior densities for undercount led to inflated medians and very wide posterior distributions of the estimated migration flows. This was especially noticeable for countries with reliable population registers, such as Sweden, Norway and the Netherlands.

As a result of our assessment, we considered a different specification for the undercount parameters. Rather than making a distinction between intra-European flows and flows to and from the rest of the world, an expert within our project grouped the countries into two categories: low and high undercount. The opinions for this new specification were also provided by this person. The answers in terms of P_1 and P_2 in Equation (3) were as follows:

- Low undercount countries: The Netherlands, Sweden, Finland, Norway, Denmark, Germany, Iceland, Austria, Belgium, United Kingdom, Cyprus, Ireland, Italy, France, Luxembourg, Switzerland, and immigration to Spain.
  - Emigration: undercount of 20–30% with 60% certainty.
  - Immigration: undercount of 5–15% with 75% certainty.
- High undercount countries: Bulgaria, Estonia, Lithuania, Latvia, Poland, Slovenia, Slovakia, Romania, the Czech Republic, Greece, Hungary, Liechtenstein, Malta, Portugal, and emigration from Spain.
  - Emigration: undercount of 50–60% with 60% certainty.
  - Immigration: undercount of 25–35% with 60% certainty.

This information was then used to construct the prior densities in the same way as described in Subsection 4.1 and resulted in posterior distributions reflecting the assessed differences in the quality of the available data.

We also investigated whether expert opinion on undercount could be removed from the model in two ways. First, we replaced the expert-based prior densities with noninformative uniform prior densities for parameters constrained between zero and one. While we were able to obtain some information concerning the differences between the high category and low category undercount, the level could not be determined purely from the data. Second, we replaced the expert-based prior densities with the noninformative prior densities and assumed all countries had the same level of undercount. In this case, the estimation algorithm did not converge.

The expert-based duration of stay prior densities were examined by keeping the constraints in Equation (10) the same and assuming weakly informative prior densities for the duration parameters in the model described in (Raymer et al. 2013). As it was mentioned in Subsection 4.2, information about the 'no time limit' and 'permanent' criteria was elicited from participants in a workshop organised by the authors. The answers were then transformed into densities following the method outlined in Subsection 4.2. We found that the outcomes were moderately sensitive to the prior densities for the duration of stay parameters. In particular, for the countries with no time limit criterion, the estimated migration flows were lower by only 6–9%, for the three-month criterion, the model with weakly informative prior densities yielded slightly larger estimates (by 4–5%), whereas for the six-month, twelve-month and permanent duration, the differences were smaller than 2%. For individual flows between countries, the differences were seldom larger than $\pm 5\%$, except for countries not providing data for flows from or to the rest of the world. Here, the differences oscillated around ± 10–15%. Finally, the uncertainty of the flow estimates was unaffected by using weakly informative prior densities.

To assess the sensitivity of the results to the expert-based prior densities for accuracy, we analysed the model in (Raymer et al. 2013) using weakly informative prior densities. The classification of accuracies of the data collection systems in countries remained the same as described in Section 2. In general, this sensitivity analysis showed that the expert-based prior densities, which reflected lack of consensus among experts about accuracy of the data collection, produced nearly the same patterns as when weakly informative prior densities were assumed. This outcome confirms the difficulty of assessing the accuracy of data collection systems.

6. Lessons Learned

As was mentioned in the literature review, elicitation of subjective opinions is a difficult task. Hence, retrospective reflections on the process as well as lessons learned during it can be as valuable as the results themselves. What did this project teach us about elicitation of expert opinion? We mention four points.

First, in our initial analyses of undercount we found that the results are sensitive to the way we specified prior densities, as reported in Section 5. The reason for this problem is not entirely clear. One explanation could be that there is very little information about migration flows to and from Europe, and experts were very uncertain about the undercount, much more so than for intra-European flows. The fact that we found stable results by reformulating the model and distinguishing between two broad categories of countries (rather than distinguishing between intra-European flows and flows to and from the rest of the world) gives some support to this explanation. Therefore, a general lesson is that it may be useful to combine extremely uncertain parameters with ones that are more certain.

Second, the notion of 'undercount of migration flows' expressed as a percentage turned out to have different meanings for different experts. In the first round one of the questions was *By how many per cent do you expect that emigration flows are undercounted in the published statistics, as compared to the true total level of emigration? Please provide a range in percentages*. The idea was that an undercount of 40%, say, reflects a situation

where the published number is 40% lower than the real (unknown) flow. But some experts gave answers that exceeded 100%. We contacted them to verify that their interpretation of an undercount of 200%, say, was as follows: The true flow is three times as large as the reported flow. In Round 2, we improved the wording of the questions on undercount. This example shows that our pilot survey was too limited (two team members and two external experts). Moreover, the testing round could have included various formulations of questions about probabilities (odds, probability, percentage or real example), which would allow us and the experts to check their consistency.

Third, the formulation of questions lacked information about the complement of the range provided by the expert. For the undercount, we did not explain to the experts that the complement of the certainty c, that is $1 - c$, is distributed to the values of the undercount outside the specified interval (but inside the interval [0, 1]). Hence, the probability mass expressed in terms of c lacked context (Gigerenzer 1994; O'Hagan et al. 2006). On the other hand, we did not want to overwhelm the experts with too detailed questions. One option here could have been to ask for a judgement, such as *During last 10 years, how many times did the reported statistics fall into the specified interval?*, rather than confidence. This question would violate the assumption of exchangeability of events (as measurement in a given year is unique) but would provide a context for experts and possibly a clearer interpretation of certainty.

A fourth general lesson is that one should be careful in selecting the experts, in particular when it comes to experience with and knowledge of probabilities and uncertainty. Indeed, we had considerable problems (fortunately in the pilot survey) to convince the experts that subjective probabilities are useful information for our assessment of migration flows. During the first and the second Delphi rounds we were in close contact with two more experts who appeared to be sceptical of the task. Some of these problems might have been avoided had we included in our introductory letter a clear explanation of the two types of uncertainty: epistemic uncertainty (lack of knowledge) and aleatory uncertainty (randomness); see Jenkinson (2005). We should have also emphasised the importance of the explanations and views behind experts' judgements.

7. Conclusion

In situations where data are inconsistent and weak, the inclusion of expert judgements is essential for improving the estimation and for reflecting uncertainty. In our research on modelling migration flows (see Raymer et al. 2013 and http://www.imem.cpc.ac.uk), we sought to provide the best possible estimates and measures of uncertainty based on available data, covariate information and expert judgements. These three pieces of information subsequently can be integrated into a single model for providing harmonised estimates of migration flows amongst 31 countries in the EU and EFTA from 2002 to 2008.

In this article, we have described our methodology for obtaining expert information on migration data to supplement reported flows and covariate information. Our implementation of this methodology was the first attempt at eliciting and quantifying opinions on various aspects of the migration data collection systems. As a result, we obtained a valuable assessment of the data on migration flows. From the varying opinions on the undercount, we can conclude that the data collection systems are expected to

capture about a half of emigrants in Europe and around 60–90% of immigrants. We learned about the likely effects of different duration of stay criteria used to record migration flows, for example, the differences in reported figures between a six-month definition and twelve-month definition. Finally, the largest ambiguity concerns the assessment of the accuracy. The only conclusion that can be drawn in that respect is that the experts expect immigration to be measured with greater precision than emigration.

After two rounds of the Delphi survey, we found that experts often disagreed on the various measurement aspects of migration. The feedback from the first round did not lead to significant changes in their opinions. However, we did not aim at convergence, as this could lead to an artificial reduction of uncertainty. Moreover, we believe that due to the heterogeneity of expert judgements expressed in the survey, the results are an important assessment of the problematic quality of the data collection systems across Europe. Nonetheless, elicitation and quantification of the expert knowledge on the data collection mechanisms in Europe is desired, especially in the context set out by the Regulation (EC) No. 862/2007 of the European Parliament and of the Council of July 11, 2007. According to the Regulation, countries in the EU are required to provide statistics on migration based on the harmonised definition of a migrant to Eurostat. The Regulation allows for use of well-documented scientific estimation and modelling methods to compile statistics on migration. Expert knowledge expressed in terms of probability distributions, as described in this article, can provide an important input to models for harmonising migration data. It also helps to understand the data collection mechanisms applied in Europe and the differences among them, as well as to assess the quality of the data produced.

8. 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, 797–825. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2009.00636.x

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, 173, 775–796. DOI: http://www.dx.doi.org/10.1111/j.1467.985x.2009.00635.x

Clemen, R.T. and Winkler, R.L. (1990). Unanimity and Compromise Among Probability Forecasters. Management Science, 36, 767–779.

Cooke, R.M. (1991). Experts in Uncertainty: Opinion and Subjective Probability in Science. New York: Oxford University Press.

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). Overcoming the Problems of Inconsistent 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.

Dey, D.K. and Liu, J. (2007). A Quantitative Study of Quantile Based Direct Prior Elicitation from Expert Opinion. Bayesian Analysis, 2, 137–166. DOI: http://www.dx.doi.org/10.1214/07-BA206

Drbohlav, D. (1996). The Probable Future of European East-West International Migration-Selected Aspects. In Central Europe after the Fall of the Iron Curtain; Geopolitical Perspectives, Spatial Patterns and Trends, F.W. Carter, P. Jordan, and V. Rey (eds). Frankfurt: Lang, 269–296.

Garthwaite, P., Kadane, J.B., and O'Hagan, A. (2005). Statistical Methods for Eliciting Proability Distributions. Journal of the American Statistical Association, 100, 680–700. DOI: http://www.dx.doi.org/10.1198/016214505000000105

Gigerenzer, G. (1994). Why the Distinction Between the Single Event Probabilities and Frequencies is Important for Psychology (and vice-versa). In Subjective Probability, G. Wright, P. Ayton (eds). Chichester: John Wiley, 129–161.

Goodwin, P. and Wright, G. (1998). Decision Analysis For Management Judgement (2nd Edition). Chichester: John Wiley.

Jandl, M., Hollomey, C., and Stepien, A. (2007). Migration and Irregular Work in Austria. Results of a Delphi-Study. International Migration Papers 90. International Labour Office; International Centre for Migration Policy Development. Geneva: ILO.

Jenkinson, D. (2005). The Elicitation of Probabilities: A Review of the Statistical Literature. Department of Probability and Statistics, University of Sheffield, Sheffield UK.

Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translation. Biometrica, 36, 149–176.

Kadane, J.B. and Wolfson, L.J. (1998). Experiences in Elicitation. The Statistician, 47, 3–19. DOI: http://www.dx.doi.org/10.1111/1467-9884.00113

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. Available at: http://mimosa.gedap.be/Documents/Mimosa_2009.pdf (accessed November 2012).

Nowok, B. (2010). Harmonization by Simulation: A Contribution to Comparable International Migration Statistics in Europe. Amsterdam: Rozenberg Publishers.

Nowok, B. and Willekens, F. (2011). A Probabilistic Framework for Harmonisation of Migration Statistics. Population, Space and Place, 17, 521–533. DOI: http://www.dx.doi.org/10.1002/psp.624

O'Hagan, A. (1998). Eliciting Expert Beliefs in Substantial Practical Applications. The Statistician, 47, 21–35. DOI: http://www.dx.doi.org/10.1111/1467-9884.00114

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., and Rakow, T. (2006). Uncertain Judgements: Eliciting Experts Probabilities. New York: Wiley.

Poulain, M. (1993). Confrontation des Statistiques de Migrations Intra-Européennes: Vers Plus D'harmonisation? European Journal of Population, 9, 353–381.

Poulain, M. and Dal, L. (2008). Estimation of Flows within the Intra-EU Migration Matrix. Report for the MIMOSA project. Available at: http://mimosa.gedap.be/Documents/Poulain_2008.pdf (accessed November 2012).

Poulain, M., Perrin, N., and Singleton, A. (Eds) (2006). THESIM: Towards Harmonised European Statistics on International Migration. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Raymer, J., Wiśniowski, A., Forster, J., Smith, P.W.F., and Bijak, J. (2013). Integrated Modeling of European Migration. Journal of the American Statistical Association, 108, 801–819. DOI: http://www.dx.doi.org/10.1080/01621459.2013.789435

Rowe, G. and Wright, G. (1999). The Delphi Technique as a Forecasting Tool: Issues and Analysis. International Journal of Forecasting, 15, 353–375.

Rowe, G. and Wright, G. (2001). Experts Opinions in Forecasting: The Role of the Delphi Technique. In Principles of Forecasting: A Handbook of Researchers and Practitioners, J.S. Armstrong (ed.). Boston: Kluwer Academic Publications, 125–144.

Szreder, M. and Osiewalski J. (1992). Subjective Probability Distributions in Bayesian Estimation of All-Excess-Demand Models. Discussion paper in Economics, 92-7. University of Leicester, Leicester.

United Nations (1998). Recommendations on Statistics of International Migration. Statistical Papers Series M, No. 58, Revision 1. New York: Department of Economic and Social Affairs, Statistics Division, United Nations.

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. Available at: http://mimosa.gedap.be/Documents/Erf_2007.pdf (accessed November 2012).

Wiśniowski, A. and J. Bijak, J. (2009). Elicitation of Expert Knowledge for Migration Forecasts Using a Delphi Survey, CEFMR Working Paper, 2/2009. Warsaw: Central European Forum for Migration and Population Research.

Received July 2012
Revised May 2013
Accepted June 2013