UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN UNION

Joint UNECE/Eurostat Work Session on Migration Statistics,
(Geneva, Switzerland, 14-16 April 2010)

Working paper 14
7 April 2010

Topic 6 of the provisional agenda

REPORTS ON VARIOUS ACTIVITIES AND PROGRAMMES

The IMEM model for estimating international migration flows in the European Union

Note by the MIMOSA*

I. INTRODUCTION

IMEM (Integrated Modelling of European Migration) is a two-year project funded by NORFACE (New Opportunities for Research Funding Agency Co-operation in Europe) to develop an integrated model for estimating migration flows between countries in Europe. This project also involves researchers from the Netherlands Interdisciplinary Demographic Institute (NIDI) and the University of Oslo.

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

The advantages in having a consistent and reliable set of migration flows are numerous. Estimates of migration flows are needed so that governments have the means to improve their planning policies directed at supplying particular social services or at influencing levels of migration. This is important because migration is currently (and increasingly) the major factor contributing to population change. Furthermore, our understanding of how or why populations change requires reliable information about migrants. Without this, the ability to predict, control or understand that

* Prepared by James Raymer, Jonathan J. Forster, Peter W.F. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy Abel, Southampton Statistical Sciences Research Institute, University of Southampton

change is limited. Finally, countries are now required to provide harmonised migration flow statistics to Eurostat as part of a new regulation passed by the European Parliament. Recognising the many obstacles with existing data, Article 9 of the Regulation states that 'As part of the statistics process, scientifically based and well documented statistical estimation methods may be used.'[1] Our proposed framework helps countries achieve this aim and provides measures of accuracy required for understanding the estimated parameters and flows.

II. BACKGROUND

The reasons for international migration are many. People move for employment, family reunion or amenity reasons. Reported statistics on these flows, on the other hand, are relatively confusing or nonexistent. There are two main reasons. First, no consensus exists on what exactly is a "migration". Therefore, comparative analyses suffer from differing national views concerning who is a migrant. Second, the event of migration is rarely measured directly. Often it is inferred by a comparison of places of residence at two points in time or as a change in residence recorded by a population registration system. The challenge is compounded because countries use different methods of data collection. Migration statistics may come from administrative data, decennial population censuses or surveys.

The timing criterion used to identify international migrants varies considerably between countries. For population register data, international migration may refer to persons who plan to live or have lived in a different country for a minimum period of three months, six months, one year, or even more. Research is needed to reconcile the different timings used to collect or model migration data, as well as between different collection systems.

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

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

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

migration data and provided a simple mathematical method for adjusting the flows and making them more consistent across countries. The multiplicative component approach showed how standard spatial interaction models for internal migration could be applied to model international migration flows in a hierarchical manner. Finally, the Bayesian approach demonstrated the usefulness and flexibility of incorporating various forms of prior information and the importance of distributions quantifying uncertainty in the predicted values.

Recently, Raymer with colleagues at NIDI have collaborated on a Eurostat-funded project to estimate international migration stocks and flows in Europe. The work on estimating flows is described in Raymer and Abel (2008). The methodology adopted by the MIMOSA (MIgration MOdelling for Statistical Analyses) team represents a two-stage hierarchical procedure. The first stage harmonises the available data by using a simple optimisation procedure (Poulain 1999) benchmarked to Sweden's migration flow data, which are assumed to be measured more or less without error (see also de Beer et al. 2009). The second stage estimates the missing marginal data and associations between countries by using the available flows and covariate information. Both stages are set within a multiplicative framework for analysing migration flows. No measures of uncertainty are provided and the approach is sensitive to the model assumptions and estimation procedure.

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

III. METHODOLOGY

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

The Statistical Modelling Framework

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

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

We first briefly describe the model of Brierley *et al*. (2008), which is used as the starting-point for the proposed research. The relationship between the true flows **Y** and reported flows **Z** may be modelled as

$$\log z_{ij} \sim N\left(\log y_{ij}, \tau^2\right), \qquad i \neq j, \tag{1}$$

where, in this case, the $\log z_{ij}$ are independent and flows are not constrained to be integers.

The Bayesian approach requires a *prior distribution* reflecting prior belief about the unknown quantities. Following previous work of Raymer (2007), Brierley *et al*. specified that the true migration flows (on a logarithmic scale) followed a prior distribution which was centred on a model similar to quasi-independence, so that

$$\log y_{ij} \sim N\left(\mu + \alpha_i + \beta_j, \sigma^2\right). \tag{2}$$

The main idea behind this model was to produce estimates that closely corresponded with the observed values, given that they must also satisfy margins which were assumed known and that quasi-independence fitted values may provide reasonable estimates of the true migration flows. The prior is particularly important for the estimation of missing flows, as it facilitates the ‘borrowing of strength’ across observed flows which may contribute information to the estimation of missing flows. This model is clearly only a starting point. It makes unrealistic assumptions and does not incorporate covariate information on migration. Note that Brierley *et al*. introduced conventional vague prior distributions for the model parameters. Because the model is complex, they were required to develop a Markov Chain Monte Carlo method to compute the posterior distribution of the migration flow matrix **Y**, and hence, to provide estimates of flows between ten Northern European nations, and associated measures of uncertainty.

Based on the above work and experience of the investigators involved in the more recent MIMOSA project, we believe that a statistical framework for migration estimation should be fully integrated, and capable of incorporating multiple data sources and expert opinion and of producing reliable estimates with levels of uncertainty. This is only achievable by developing a Bayesian methodology. In particular, we are developing a methodology that substantially extends the exploratory log-linear model developed by Brierley *et al*. (2008) in the following ways.

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

$$z_{ijt}^{S} \sim Po\left(\mu_{ijt}^{S}\right), \qquad (1)$$
$$z_{ijt}^{R} \sim Po\left(\mu_{ijt}^{R}\right). \qquad (2)$$

In our model, y_{ijt} is a true flow of migration from country *i* to country *j* in year *t*. In terms of measurement, true flows are consistent with the United Nations (1998) recommendation for long-term international migration. The two measurement error equations are

$$\log \mu_{ijt}^{S} = \log y_{ijt} + \beta_i + l(\kappa_i) + \varepsilon_{ijt}^{S}, \qquad (3)$$
$$\log \mu_{ijt}^{R} = \log y_{ijt} + \gamma_j + l(\kappa_j) + \varepsilon_{ijt}^{R}, \qquad (4)$$

where we assume $\varepsilon_{ijt}^{S} \sim N(0,\tau^{S})$ and $\varepsilon_{ijt}^{R} \sim N(0,\tau^{R})$. The variances of the error term do not depend on the country. Instead, they simply depend on whether the data are captured by sending or receiving countries. It is planned to extend this to allow them to vary by type of collection system (e.g., population register or survey). The number of parameters required to capture differences in accuracy will ultimately depend on our typology of collection systems and their relative ability to capture migration flows, regardless of definition and coverage.

Differences in duration and undercount of emigration data are captured by the parameters $\beta_i = \delta_{def(i)} - \lambda$ and $\gamma_j = \delta_{def(j)}$, where δ_m is the effect of using duration definition *m*, *def*(*i*) is the definition used be country i and λ is the effect of the undercount. Note that the δ for countries using the UN definition for duration is zero. The country-specific parameter κ captures the differences in coverage with respect to the UN (1998) definition of migration. A log-inverse-logistic transformation of κ, $l(\kappa_i) = -\log[1+\exp(-\kappa_i)]$, is included in the models. This allows us to set a normal prior for κ.

The true flows of migration may be modelled according to a set of covariate information.
Here, we rely on migration theory and empirical evidence to drive the development of the model. The following sets out an initial model that will later be extended to include further variables. To begin, consider the following 'gravity' model of migration:

$$\log y_{ij} = \alpha_1 + \alpha_2 \log(P_i) + \alpha_3 \log(P_j) + \alpha_4 d_{ij} + \xi_{ij}, \qquad (5)$$

where P_i and P_j are the mid-year populations in sending and receiving country, respectively, d_{ij} is a dummy variable measuring contiguity (or neighbouring countries) equalling one if countries *i* and *j* have a common border and zero otherwise, and $\alpha = (\alpha_1,...,\alpha_4)^T$ is a vector of structural parameters. Here α_2 and α_3 are elasticities of *y* with respect to population size. A priori, it is expected that they are positive and equal to one (i.e., they capture the proportional relative change of migration with respect to relative change in population). Also, it is expected that α_1 is negative (average migration is a fraction of the population size) and α_4 is positive (i.e. migration between bordering countries is larger). The random term ξ is assumed to be normally distributed with mean zero and constant variance σ^2.

We plan to use covariates similar to those used in the MIMOSA project and by Abel (forthcoming). Some issues may arise regarding the specification of priors for these variables. The choice of covariates and the specification of the prior distributions are important as this model will provide the only information for estimating the flows in the countries where the data are missing.

For the moment, we assume that all the parameters are constant over time. However, later on, we plan to test whether this assumption is valid. For instance, we may want to include some time-structure over the parameter space.

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

$$f\left(y, z, \mu, \delta, \lambda, \kappa, \tau, \alpha, \sigma^2 \mid x\right) = f\left(z \mid \mu\right)\, f\left(\mu \mid y, \lambda, \delta, \kappa, \tau\right)\, f\left(y \mid \alpha, \sigma^2; x\right) \\ f(\lambda)\, f(\delta)\, f(\kappa)\, f(\tau)\, f(\alpha)\, f\left(\sigma^2\right), \qquad (6)$$

where $f(z \mid \mu)$ is the data model, $f(\mu \mid y, \lambda, \delta, \kappa, \tau)$ is the measurement model, $f(y \mid \alpha, \sigma^2; x)$ is the migration model and $f(\lambda)$, $f(\delta)$, $f(\kappa)$, $f(\tau)$, $f(\alpha)$ and $f(\sigma^2)$ are the priors. The distribution of *y* given the observed flows, *z*, can be obtained by integrating every other parameter out of density (6). Note, this characterises a Bayesian approach.

Constructing the Prior Distributions

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

The elicitation of prior information involves specifying the quality of data sources, the differences in definitions and the role of the explanatory variables. Some information has to be elicited from external experts, other information will be provided by team members. Experts will be asked to rate the credibility they give to data arising from different types of data collection mechanism (e.g., survey versus register) and emigration versus immigration. Further, we will ask the experts about rates of bias (e.g., systematic undercount) in some (or all) flows. Each expert will be asked to supply a set of distributions representing their beliefs about certain model parameters. However, it is likely that experts will only be able to provide summary statistics concerning a subset of these parameters. The totality of expert opinions will need to be combined into a single set of distributions, allowing for the introduction of yet another source of uncertainty, related to the heterogeneity of experts. To keep it under control, a multi-stage process of extraction (elicitation) of expert judgement will be used, for example within a Delphi survey, whereby the expert opinions are allowed to converge towards a common consensus. For a discussion of the approach and potential problems related to the elicitation of judgement specifically from migration experts, see Bijak and Wiśniowski (forthcoming).

Another critical aspect of prior elicitation will be to obtain information concerning which covariate information should be included in the model (e.g., relative GDP, relative unemployment, language, contiguity, etc.) and the size of any associated model parameters. This information will be based on a meta analysis of existing studies. Experts will also be asked to provide their relevant knowledge concerning covariates. Here, this project builds on the IDEA project, run within the EC 6th Framework Programme in 2007-2009. IDEA concerns how certain European countries have changed from net senders of migrants to net receivers of migrants and includes the preparation of migration forecasts and policy recommendations. Bijak and Wiśniowski (forthcoming) provide a good basis for including elicitation of expert knowledge for use in probabilistic (Bayesian) migration modelling. The research further develops these ideas, in the specific context of the multi-country model for migration flows.

Finally, the IMEM project also builds on the experiences of the MIMOSA project which was executed on behalf of the European Commission during the period 2007-2009, coordinated by NIDI and involving Raymer, van der Erf, Abel, Bijak and Wiśniowski. The aim of MIMOSA project was to develop appropriate methodologies to reconcile the differences in international migration statistics in European countries and to provide consultancy to both Eurostat and Member States of the European Union (EU) on the way to produce more reliable migration figures. The results of MIMOSA (without measures of uncertainty) will serve as a useful starting point for the current project and a basis for comparing the results.

IV. DATA COLLECTION

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

To obtain data on expert judgements, we will build on the experiences gathered in the frameworks of the IDEA project and the MIMOSA project. The form in which judgemental data will be elicited from the experts will be determined at the initial stage of the project. We will choose a method and tools that allow for a proper translation of subjective expert knowledge into probability distributions to be used in the Bayesian modelling. The elicitation method will first be subject to internal testing by the research teams, so as to ensure that it provides unambiguous outcomes in line with the research needs. External experts will be selected based on their familiarity with migration statistics and statistical thinking and will be invited to be involved in one or more prior elicitation activities, including a workshop discussion, a one-off survey and a multi-stage (Delphi) survey. The information they provide to the team consists of both an assessment of the quality of the data sources and an assessment of the importance of explanatory variables.

Subsequently, the aggregated judgements of the experts will be translated into prior probability distributions to be used as input in Bayesian models. At this stage, various formulations of expert knowledge will be taken into consideration. We will then compare the outcomes yielded by different possibilities, as well as by the 'reference' (non- or weakly-informative) prior distribution. In this way, we will be able to assess how robust our results are against technicalities in the formalisation of judgemental information, and how important expert judgement is in our study.

V. CONCLUSION

This paper has presented the initial model framework of the IMEM project. Prototype testing has been done on a subset of countries and it appears to be a promising approach. The models are being programmed in both R and WinBUGS. The next steps for the project are to continue developing the model, including the borrowing of strength over time, and to elicit expert information on the definition, coverage and accuracy aspects of the model.

Literature List

Abel GJ (forthcoming) Estimation of international migration flow tables in Europe. *Journal of the Royal Statistical Society, Series A*.

Bijak J, Wiśniowski A (forthcoming) Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society, Series A*, 173 (4).

Brierley MJ, Forster JJ, McDonald JW and Smith PWF (2008) Bayesian estimation of migration flows. In *International migration in Europe: Data, models and estimates*, Raymer J and Willekens F, eds., pp. 149-174. Chichester: Wiley.

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

de Beer J, van der Erf R and Raymer J (2009) Estimates of OD matrix by broad group of citizenship, sex and age, 2002-2007. MIMOSA Deliverable 9.2 B, Netherlands Interdisciplinary Demographic Institute, The Hague.

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

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

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

Girosi F and King G (2008) *Demographic Forecasting*. Princeton: Princeton University Press.

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

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

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

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

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

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

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

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*, Raymer J and Willekens F, eds., pp. 209-234. Chichester: Wiley.

Raymer J and Abel G (2008) The MIMOSA model for estimating international migration flows in the European Union. Joint UNECE/Eurostat Work Session on Migration Statistics, Working paper 8. Geneva: UNECE/Eurostat.

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

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

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 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 JM (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.

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