Integrated Modelling of European Migration Database Case Study

Arkadiusz Wiśniowski, Martin Chivers and Michael Whitton

1. Introduction

The Integrated Modelling of European Migration (IMEM) Database is an on-line tool for presenting the estimated international migration flows amongst countries in the European Union (EU) and European Free Trade Association (EFTA), as well as to and from the rest of the world, in the period 2002 to 2008. It allows extraction of various information and characteristics of the estimated flows, such as country of origin, country of destination, age and sex of migrants. The database can be used by academics, official statisticians or policymakers, who would like to obtain information about the recent migration dynamics in Europe.

The estimated international migration flows are the output of the project Integrated Modelling of European Migration (IMEM), which was funded by New Opportunities for Research Funding Agency Cooperation in Europe (NORFACE). The collaborating parties came from Southampton Statistical Sciences Research Institute (S3RI), University of Oslo and Netherlands Interdisciplinary Demographic Institute. It is well acknowledged that the official statistics on migration flows suffer from underreporting, are incomparable between countries or are unavailable. The IMEM project sought to provide a statistical framework for modelling migration flows amongst countries Europe in the context of inconsistent, inadequate and missing data. The ultimate output was a table of migration flows harmonised to a common definition.

2. Personnel and background

Initially a meeting was held to investigate how the IMEM project and DataPool could work together for a disciplinary exemplar. This included the Principal Investigators for both projects, James Raymer and Wendy White respectively. It was agreed that creating a database tool to encourage wider use and impact would be of most use to the IMEM.

Arkadiusz Wiśniowski (Research Fellow on the project IMEM at S3RI, currently at the ESRC Research Centre for Population Change, University of Southampton) and Michael Whitton (Academic Liaison Librarian) were tasked to work on the exemplar. Since IT development would be a critical part, representation from iSolutions was requested and received. 'Off the shelf' solutions were investigated, including Nesstar[1] (used by the UK Data Archive). However, they were not found to fit the requirements – being unable to deal with types of queries (mean, quartiles, etc.) needed.

iSolutions began some initial scoping work. However, the development of the tool took longer than desired. The technical aspects required specific skills of a database and web development specialist. There was limited staff fitting this profile (busy with critical projects) which caused a delay. Once Martin Chivers (Database & Web Developer, iSolutions) was assigned to the project development progressed quickly.

Becki Davies, the Knowledge Exchange Manager at the ESRC Research Centre for Population Change (CPC), was involved in designing the guidance for the database website.

The Integrated Modelling of European Migration (IMEM) Database and associated guidance is available online[2].

3. Technical description of the IMEM Database

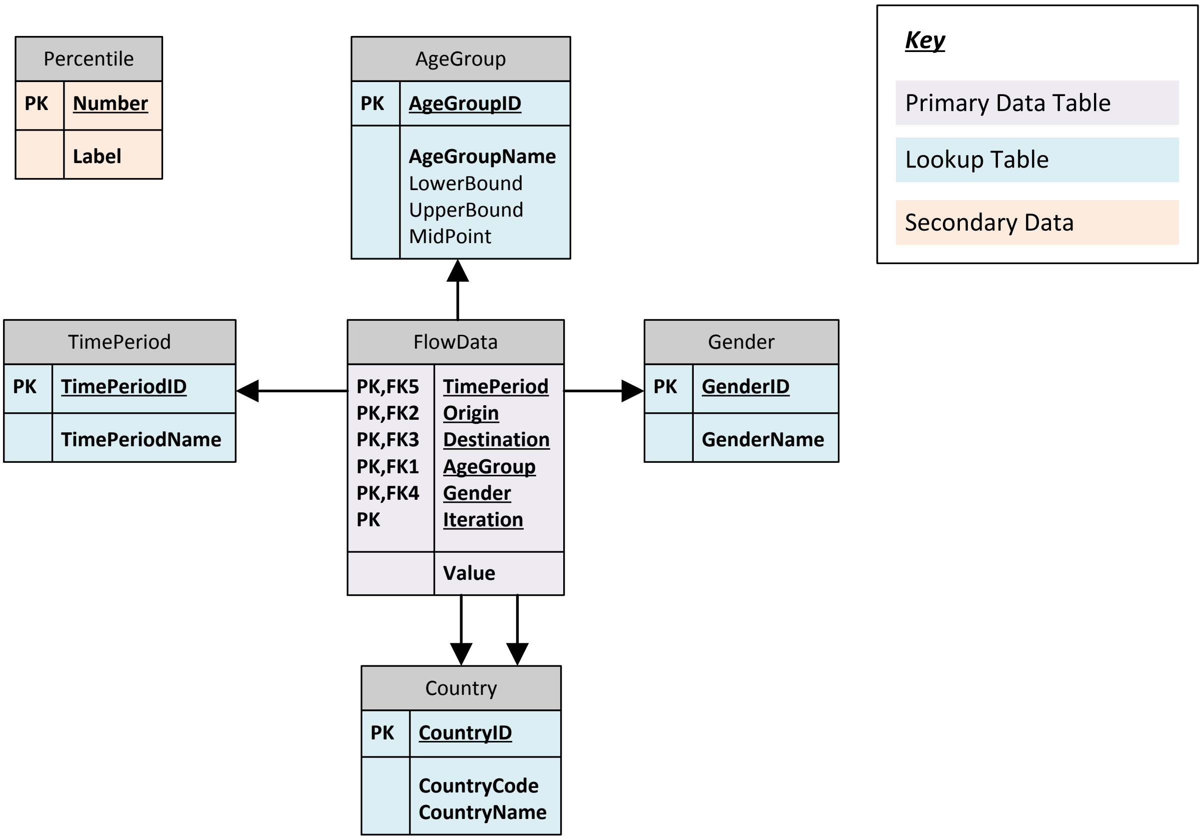
The database and web application were developed using Microsoft SQL Server 2008 (database software), and Microsoft ASP.NET (web pages). The chart generation components within the .NET Framework v4.0 (System.Web.UI.DataVisualization) were used to generate the charts.

The data were supplied to us in the form of a simple tuple (see sample below) comprising time period, origin, destination, age group, gender, iteration, and estimated flow – this had effectively been pre-categorised, as each variable was encoded to an integer, instead of using the "human readable" string representation of it, with a set of definition tables being supplied for each variable.

Time	Origin	Destination	Age	Sex	Iteration	Value
1	2	1	1	1	1	10.661
2	2	1	1	1	1	9.067
1	1	2	1	1	1	6.8304
2	1	2	1	1	1	7.139

Due to the nature of the data, the database is of a relatively simple construction – there is a single "primary" table which contains the detail of the estimated flow, with several "lookup tables" containing a human-readable description of the keys stored – this helps keep the database's footprint on disk small, and limits the amount of repetition of data. An additional table of percentiles was created, to simplify queries.

[1] http://www.nesstar.com
[2] http://www.imem.cpc.ac.uk



Whilst the number of records in the lookup tables is relatively small (at most 32 records), as the dataset is essentially the Cartesian product of the variables (albeit with the destination count decremented by 1, as "identity flows" are omitted), the primary data table is somewhat large – 249,948,000 rows. Whilst the smallest available data types were used to encode the data (typically 8-bit integers for the row keys), this still resulted in the primary table being 5.65GB (prior to indexes being added).

Due to the nature of the data, the actual rows would not be returned to the user – instead, the computed median and associated measures of uncertainty would be returned. Given that a simple query to extract the immigration to a single country would need to retrieve 7.8m rows, query performance is a definite concern. The volume of data returned also meant that it was impractical to return the full results of the query to the web application, and for the median, etc., to be calculated there. This is somewhat compounded by the lack of a median or percentile function in Microsoft SQL Server 2008, which meant both would have to be implemented either in T-SQL, or as an user-defined aggregate function from embedded .NET assembly. Ultimately, the median function was implemented in T-SQL, based upon code written by Itzik Ben-Gan[3], but generalised so that any

[3] http://www.sqlperformance.com/2012/08/t-sql-queries/median

percentile could be returned, rather than just the 50th (thus enabling all the measures of uncertainty to be returned in a single set-based operation).

One “interesting” problem was found when executing the SQL code – for a given instance of a query, the database engine would generate substantially different query execution plans for a parameterised version of the query, when compared to one where the parameters were bound into the query in literal form, despite the actual query being identical. Unfortunately, this caused a difference in query execution speed of approximately one order of magnitude. Therefore, the application was altered to incorporate the parameters as literals within the query, rather than as parameters. This is generally not the accepted practice, as it creates one execution plan for each query (thus leading to other potential performance issues, and polluting the execution plan pool). However, given the performance difference, it was the only option. The resultant query was still passed through sp_executesql, to ensure that the execution plan was calculated once and cached for future executions, so the plan did not need to be recalculated later.

The final point to be addressed was how to display the data. Were the application displaying “actual” migration data (rather than estimated data), then a highly-graphical method would have been preferred – for example, taking a map of the EU member states, and overlaying arrows between them to indicate the flows present for the selected query options, with the size of the arrow denoting the amount of movement (possibly relative to the largest flow returned, rather than the absolute flow size). However, due to the data being returned being estimates, the above approach may not be appropriate, as it may indicate a false confidence in the numbers (in essence, there is no simple way of illustrating the measures of uncertainty). Therefore, it was decided that line charts should be employed, including the median bracketed by percentile ranges.

4. Impact of the project

Potential users of the IMEM database include: policymakers of various levels of authority, academics and researchers in social sciences, national statistical institutions, NGOs and students. The database provides a long-term and sustainable resource for future use, as well as potential for expansion of the range of the available statistics. There is a commitment within the Institution to maintain the platform where the service is hosted, and to upgrade to appropriate software.

There are four main contributions of having an on-line database with the estimated migration flows in Europe. First, it permits access to the entire range of the estimates of the harmonised migration flows for 31 countries in the EU and EFTA. Different measurements of the flows can result in very different patterns. The IMEM estimates directly account for the main differences found in the measurement aspects of the reported data on migration. Before the IMEM study, little was known about the effects of measurement, and no one had attempted to model the differences by considering the main aspects of defining and measuring migration flows: various duration criteria used by countries, underreporting of migrants, coverage of subpopulations and accuracy of the measurement. The estimated flows are consistent with the United Nations recommendation for the measurement of international migration.

This work is especially relevant when considering the expansions of the EU in 2004 and 2007. The database provides sound evidence about recent migration patterns in Europe and can be used for informed policymaking. One of the conclusions of the IMEM project is a need for the cross-national exchange of information on migration. The database is a medium through which the results can be communicated to the official statisticians, who are interested in comparing their own figures on migration with the estimates that are produced when their figures are combined with other countries' data.

Second, the IMEM Database can be used by the academics and researchers to study reasons and implications of migration, especially before and after the EU enlargement in 2004 and 2007. Various characteristics of the estimates that are accessible from the IMEM Database can be used to test theories in economics, demography or sociology. Moreover, through the international visibility and accessibility of the database, new research networks and collaborations can be stimulated. An interest has already been expressed by the researchers involved in the forecasting part of the project funded by ESPON: *Territorial Scenarios and Visions for Europe ET2050*. However, at the time of the enquiry (September 2012), the database was not complete.

Third, the on-line database permits a clear presentation of the measures of uncertainty for the estimated flows in the form of quantiles of the probability distributions. These measures are relevant for assessing the quality of the reported flows and of the estimates. In the area of combining data from different sources and missing data, it is important to be clear about the accuracy of the estimated figures.

Fourth, the database is an example on how the results of the statistical model estimated by using Bayesian approach can be presented. Samples from the probability distributions are the output of the Bayesian statistical model. Presentation of full distributions requires using more advanced tools than the ones applied to create the IMEM database. Furthermore, interpretation of the results presented in the form of distributions would require the potential users to have some training in statistics. Hence, the results are currently presented in a user-friendly manner.

Impact of the IMEM Database can be measured by:

- Citations of using the results in the publications,
- Creation of databases, the design of which replicates the design of the IMEM Database.

5. Lessons learned

Creation of the IMEM database required collaboration of a person highly skilled in database design (Martin Chivers, iSolutions) and a person engaged in the process of the estimation of the migration flows in the IMEM project (Arkadiusz Wiśniowski, S3RI). The particular requirements and communication between the parties was ensured by Michael Whitton (Academic Liaison Librarian, Hartley Library).

The process of creation of the IMEM database can be split into four phases:

1. Description of the data and required characteristics to be accessible for the end user,
2. Preparation of the data,
3. Design of the tool for extraction of the desired characteristics of migration flows,
4. Design of the contents of the Internet website for the end users of the database.

In the first phase the most important task was to bring together the person with the skills of database design and the supplier of the data to be used in the database, and to ensure good communication between them. This was ensured by Michael Whitton. Further, the key points of the IMEM project, as well as the particular aspects of the final results of the project, were described to Martin Chivers by Arkadiusz Wiśniowski. At this stage, the key lessons were:

- There is a need for services rendered by iSolutions to researchers;
- A proper description of the requirements of researchers is crucial for the proper design of the final product;
- A proper description and presentation of the available resources and potential solutions is important to ensure time and cost efficiency of the undertaking on the side of the iSolutions.

The data have been prepared according to the requirements of Martin Chivers as a single CSV file produced in the MATLAB software. Since the file was large (5.65GB), it was delivered by using the University of Southampton Drop-off service. The key lesson:

- The drop-off service would be particularly required if the process involved exchange of large files on a regular or frequent basis.

Detailed technical description of the tool for extraction of the desired characteristics of migration flows is presented in Section 3 of this report. The process involved communication of the progress on a regular basis. Communication was enhanced by using the open source 'Trac' system to keep track of bugs and fixes. The key lessons of this phase are:

- Frequent exchange of information and providing examples is critical to ensure that the tool is designed as required by the researchers.
- The tracking system can greatly enhance this process as it enables multiple parties to participate in discussion and development.

The last phase of the IMEM Database creation required skills in designing a user-friendly interface of the website where the tool is hosted. For that purpose, Becki Davies (CPC) was engaged in the process. The particular issues involved preparation of the instructions on how to use the tool. To help users understand how the database operates, examples have been provided. Lesson learned in this stage:

- It is important to understand the needs of the user who has not been involved in the preparation of the results or lacks the statistical skills to interpret the results.
- Examples are simple and effective in explaining how the tool should be used.

6. Summary and recommendations

The IMEM Database case study has demonstrated the successful cooperation between the IT department of the University (iSolutions) and a group of researchers willing to communicate the results to the wide audience (IMEM team, S3RI). This cooperation was facilitated by the Academic Liaison Librarian, Michael Whitton. Advantages of having the database are numerous and long-term. Potential users include academics, national statistical institutes, policymakers and students. Presentation of the results on such a wide scale without this specific database would be impossible.

Since the results of the IMEM project consisted of a large number of samples from the probability distributions, it was required to design a tailored database with the tools that would allow production of the desired characteristics of these distributions. Despite the fact that the database is dedicated to presenting the results of the specific project, the design of it can be used to present the output of any research which can be presented in terms of large number of samples from the probability distributions.

Technical Challenges and Solutions

- *How to present the characteristics of the probability distributions in a clear, computationally- and cost-effective manner?* Particular characteristics of the distributions (here – quantiles) can be used as the final output.
- *How to extract the underlying data from the database in a manner with good performance?* Packing the records as densely as possible (i.e. the narrowest data type that can represent the data), and using optimised queries.

Organisational Challenges and Solutions

- Cultural differences: Recognise different cultures exist between research community and IT specialists in central services having different professional language, expectations and working practises. The management of a research project usually requires a different, iterative methodology than an IT infrastructure project having a more clearly pre-determined end point. Communication on a regular basis between the parties enhances the process.
- Evaluation of the exemplar – both the benefits from the specific tool, and also how useful a similar approach would be applied to other disciplines. This should lead to an understanding of the resource needed to deliver this (see below).
- Demands on database and web development skills – these are significant and this could increase in the future. If this approach is seen as important and significant numbers of projects need this kind of service – how can this be resourced? It may be desirable to derive some of this from grant income, either an amount costed into the bid for iSolutions or funding specialist posts for larger grants.
- Skills needed to support data management – both increasing skills of existing staff/roles and potential need for new roles. The latter may include a mixture of IT and academic skills for example.