UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M09/09
Methodology

Estimation Of International Migration Flow Tables In Europe

Guy J. Abel

Abstract

A methodology is developed to estimate comparable international migration flows between a set of countries. International migration flow data may be missing, reported by the sending country, reported by the receiving country or reported by both the sending and receiving countries. For the last situation, reported counts rarely

match due to differences in definitions and data collection systems. In this paper, data known to be of a reliable standard is used to create an incomplete migration flow table of harmonized values. Cells for which no reliable reported flows exist are then estimated from a negative binomial regression model fitted using the Expectation-Maximization (EM) algorithm. Finally, measures of precision for all missing cell estimates are derived using the Supplemented EM algorithm. Recent data on international migration between countries in Europe are used to illustrate the methodology. The results represent a complete table of comparable flows that can be used by regional policy makers and social scientists alike to better understand population behaviour and change.

Estimation of International Migration Flow Tables in Europe

Guy J. Abel*

February 2009

Abstract

A methodology is developed to estimate comparable international migration flows between a set of countries. International migration flow data may be missing, reported by the sending country, reported by the receiving country or reported by both the sending and receiving countries. For the last situation, reported counts rarely match due to differences in definitions and data collection systems. In this thesis, reported counts are harmonized using correction factors estimated from a constrained optimization procedure. Factors are applied to scale data known to be of a reliable standard, creating an incomplete migration flow table of harmonized values. Cells for which no reliable reported flows exist are then estimated from a negative binomial regression model fitted using the Expectation-Maximization (EM) algorithm. Covariate information for this model is drawn from international migration theory. Finally, measures of precision for all missing cell estimates are derived using the Supplemented EM algorithm. Recent data on international migration between countries in Europe are used to illustrate the methodology. The results represent a complete table of comparable flows that can be used by regional policy makers and social scientist alike to better understand population behaviour and change.

Keywords: Constrained Optimization; Flow Tables; International Migration; Migration Estimation; Negative Binomial Regression; SEM algorithm

1 Introduction

Migration flow data inform policy markers, the media and academic community to the level and direction of population movements. In any one country, reliable migration data provide a means to improve the governance of population flows and their impacts. They also allow a better understanding of the causes and consequences of people's movements. However, reliable migration data for comparisons of international population flows between a set of countries are often lacking. Reported counts are either missing, reported by the

*Correspondence Address: Division of Social Statistics, School of Social Sciences, University of Southampton, Highfield, Southampton, S017 1BJ, United Kingdom. Email: g.j.abel@soton.ac.uk

sending country, reported by the receiving country or reported by both the sending and receiving countries. For the last situation in which two sources of information are possible for one particular flow, reported counts rarely match due to differences in data collection and measurement.

Comparable migration data can aid concerned parties manage policy and understand people's movements better (Bell *et al.*, 2002). This is apparent for a number of reasons. First, comparative summaries of international migration flows become more meaningful when they are presented in a multinational context. Second, data from multiple nations can provide a more comprehensive empirical source for the testing of migration theories. Third, such analysis has the potential to provide new insights to the dynamics of migration between countries. Finally, the difference between public policies for international migration across multiple countries can be more readily studied when comparative measures exist.

In Europe, the study of international migration data is of growing importance due to the political reforms agreed by the European Parliament in 2004. These allows citizens in the European Union (EU) the right to move between, and reside freely in, member states (Kraler *et al.*, 2006). In recent decades, policy makers of the European Parliament have introduced legislation for the supply of international migration flow data. In 1976, Community Regulation No 311/76 required members to supply migration statistics annually to Eurostat (the statistical office of the EU). In 2007, Regulation No 862/07, obliged members to provide migration statistics that complies with a harmonized definition. However, despite these regulations migration flow data often lacks adequate measurements of volumes and direction, demographic completeness and comparability between nations (Kelly (1987), Salt (1993), Willekens (1994) and Nowok *et al.* (2006)).

This paper develops steps towards these ends by outlining a methodology that can be used to estimate comparable international migration flow data. This is undertaken by addressing two fundamental data problems: inconsistencies and incompleteness. In order to make observed data consistent, a constrained optimization procedure is used. This relies on the assumption that for selected flows the difference between reported counts by sending and receiving countries are fixed. Given a constraint on a data source(s), for which no adjustment is required, these differences can be minimized by estimating parameters to scale reported counts from each data provider. In order to make a table of these harmonized flows complete, the Expectation Maximization (EM) algorithm originally proposed by Dempster *et al.* (1977) is used to fit a spatial interaction model. This allows imputations for missing values to be obtained from model parameters estimated using the complete (rather than the observed) data. Finally, estimates of precision for the imputations are calculated using the Supplemented EM algorithm of Meng & Rubin (1991). This methodology is applied to a series of data for international migration flows between the 15 countries in the European Union (EU15) before the expansion that took place in May 2004.

Constrained optimization procedures have been used in many different contexts, including international migration flow data, for example Poulain (1993) and Poulain (1999). Such procedures appear to be an appropriate method to harmonize flow data. For reliable data sources the difference between reported counts appear constant across time (Kupiszewska & Nowok, 2008). In this paper, previous applications of constraint optimizations to international migration flow data are extended to consider alternative distance measures, constraint sets and generalized across a series of migration tables.

Previous methods for the estimation of missing international migration flows (Poulain (1999) or Raymer (2007)) have tended to be based on ad-hoc adjustments to existing data or interpolations from simplistic models. However, more satisfactory estimates can be obtained by specifying a more comprehensive model, that describes each flow in relation to others (Willekens, 1994). Parameter estimates for this model, which account for the incompleteness of the observed (harmonized) data, can be obtained using the EM algorithm. Extensions of this algorithm are relatively well developed and provide a number of neat statistical properties for parameter estimates and imputed values. Together, the application of these two methods, allow comparable international migration flow data to be estimated.

2 Problems of Comparability in International Migration Flow Data

The lack of comparability in international migration data can be traced to the multi-dimensional nature of migration (Goldstein, 1976). As a result, national statistical institutes have developed measures of migration solely suitable to their domestic priorities. Full reviews of the international migration flow data and their issues can be found in Kelly (1987), Willekens (1994), Nowok *et al.* (2006) and Kupiszewska & Nowok (2008). The incomparability between data sources in any time period are predominantly derived from

(a) differences in data production techniques,

(b) differences in the dissemination of data.

Each is discussed in relation to measures of migration flows by origin or destination.

2.1 Data Production Techniques

Differences in the production of migration flow statistics can be derived from distinctive data collection methods and definitional measurements used by national statistical institutes.

Data collection methods may influence the completeness and accuracy of reported migration flows (Nowok *et al.*, 2006). National statistical institutes collect migration flow data from a variety of sources. Computerized population registration systems that continuously cover the target population often provide reliable and timely statistics. Where administrative sources do not cover all or part of the target population other registers such as alien or residency permit databases are sometimes used. Some nations rely on surveys carried out during border crossings or among households inside a country. These can be more problematic. For example, in the United Kingdom the International Passenger Survey (IPS) is used to help provide international migration flow data. In order to provide sufficient detail for analysis the sample size must be very large otherwise unexpected irregularities appear for specific origin to destination flows (Perrin & Poulain, 2006b).

Migration definitions can influence the reported volume of movements. Definitions of migration flows involve a statement of duration and population coverage. The duration of time used to identify international migrants varies between countries (Kupiszewska & Nowok, 2008). For population register data, international migration may refer to persons who have lived in a different country for three months, six months, or one year. For census or survey data, the entry date of international migrants is not known, only that they lived outside the country one-year or five years prior to the census or survey date. In data sources the intended duration, rather than the actual duration is used. Under an actual duration measure, reporting of figures are delayed to allow the period used in the timing criteria to pass, whereas under a intended duration an assumption that the intended period will become the actual duration is made. Nowok *et al.* (2006) noted that some national statistical institutes measure intended duration of non-national immigrants by the period specified in the authorization to stay which may differ from the actual duration.

The coverage of difficult to measure population groups, such as asylum seekers, students and illegal residents, in migration definitions varies between data sources. Asylum seekers are generally included as migrants when granted permission to stay. Exceptions to this rule are found in some countries such as Germany and the Netherlands, where the registering of seekers occurs at an earlier stage of the asylum procedure (Erf *et al.*, 2006b). Erf (2007) noted that students moving between EU countries are often not included in international migration flow figures as they are not required to report their migration. However, in countries such as Denmark students are required to have residency permits on which migration data are based. Data on undocumented migrants should be included in migration figures according to most definitions used in European migration statistics regulations but are often missed due to collection difficulties. In the EU only Spain allows the registration of illegal migrants through a pardon system (Breem & Thierry, 2006b), allowing the capture of data on this difficult to measure population.

2.2 Data Dissemination Methods

National statistical institutes may struggle to fully disseminate detailed information on migrants such as their origin or destination. In such cases, the total flow in or out of the country is often known, resulting in a count of migrants with unknown countries of origin or destination. For some nations the size of these counts are relatively large with regard to the total migration count. For other nations this count may be small or zero. Hence, when comparing migration flows between multiple nations, the counts of movements associated with unknown origins or destinations must be considered.

Migration data may be partially or completely unavailable. Partial availability can occur for data from countries that have a domestic need to only measure certain flows. For example, in 2002 Ireland produced estimates of total movements to and from only three areas: the United Kingdom, the United States of America and the EU (Perrin, 2006). In other countries, partial completeness is caused by insufficient data collection methods. For example, the IPS carried out during border crossings into and out of the United Kingdom are unable to provide estimates for individuals origins or destinations where low volumes of movements exist (Perrin & Poulain, 2006b). For some countries no migration flow data may be produced. For members of the EU this failure appears to be random. For example, France which has a large volume of migration, does not register citizens entering or leaving the country (Breem & Thierry, 2006a). Conversely, similar sized countries, such as Italy, regularly publish migration flow data. In some years, migration flow data provided by countries to international organizations (the main source of international migration flow data for multiple nations) can appear as incomplete. This can be caused by national statistical institutes not providing, or the organizations not publishing data, despite collection procedures being in place.

3 Methodology

In this section, a general methodology that allows the estimation of international migration flow tables is described. In order to provide comparable estimates, inconsistencies and incompleteness in reported migration counts from differences in the production and dissemination, are addressed. This is undertaken in three stages

(a) correction for unknown counts,

(b) harmonize selected data,

(c) impute missing and ignored data.

Each stage is outlined in turn.

3.1 Correction for Unknown Counts

Migration data are commonly represented in square tables, with off diagonal entries containing the number of people moving from any given origin i, to any given destination j, in a single time period. The diagonal information in the migration flow table (which contains either counts of migration flows within an area or populations) are often omitted in an international context. As a single flow can be counted by national statistical institutes of both sending and receiving countries, two migration tables may be produced: one for receiving data collected at the destinations and one for sending data collected at the origin. Observations of these flows can be represented in an array m_{ijk}, where $k = 1, 2$ indicates receiving and sending flow tables respectively.

As previously discussed, international migration flow data are accompanied by a count of migrants with an unknown origins or destinations. In order to account for these unknowns and thus avoiding bias towards data sources with no unknowns, these flows can be adjusted,

$$m_{ij1} = n_{ij1} + \left(\frac{n_{ij1} n_{iu1}}{n_{i+1} - n_{iu1}} \right), \; m_{ij2} = n_{ij2} + \left(\frac{n_{ij2} n_{uj2}}{n_{+j2} - n_{uj2}} \right), \tag{1}$$

where n_{ijk} is the original observed migration flows, the index $i, j = u$ denotes the unknown count for the respective origin or destination and $i, j = +$ are the country total flows including unknowns counts.

3.2 Harmonize Selected Data

When reported sending and receiving migration data are plotted over time, selected flows demonstrate a constant difference between their values. This is illustrated on a logarithmic scale in Figure 1 for available data in the EU15. Origins, which provide the sending data, are shown on the vertical axis. Destinations, which provide the receiving data, are shown on the horizontal axis. Non-parallel lines are visible for reported flows in and out of some nations such as Great Britain, where British counts tend to be more volatile due to their quality (Kupiszewska & Nowok, 2008).

Differences in counts between nations with better quality data can be considered as fixed, where data production techniques do change over time. Thus measures of these differences represent the non-random discordance in the collection and measurement of migration flows between any two national statistical institutes.

Poulain (1993) took a similar view in his attempt to harmonize migration data, where by all reliable data were considered to be influenced by some data source specific correction factor. Under this assumption, when the correction factors are known, the equality

$$r_j m_{ij1} = s_i m_{ij2}, \tag{2}$$

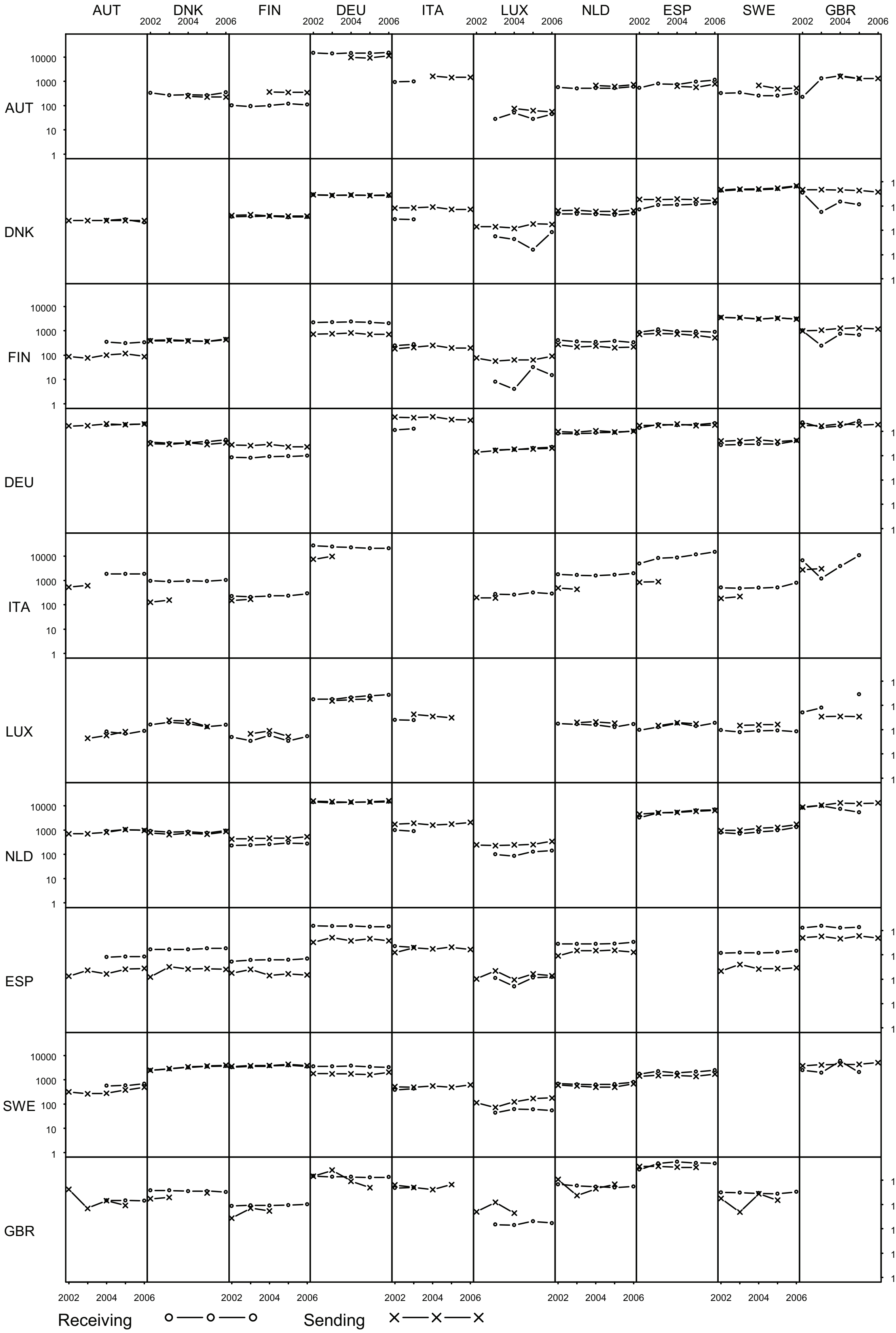

Figure 1: Logarithm of Reported Receiving and Selected Flows Between Selected EU Countries in 2002-2006.

should hold, where r_j scales receiving data and s_i scales receiving data. When they are unknown, Poulain (1993) suggested that correction factors can be estimated in each time periods by minimizing a Euclidean distance measure,

$$f(r_j, s_i|m_{ijk}) = \sum_{i,j}(r_j m_{ij1} - s_i m_{ij2})^2, \tag{3}$$

and imposing a constraint on the overall table total based on observed data. The method of Lagrange multipliers was used to estimate these unknown parameters. For his selected data the estimated values were relatively stable across time. Alternative distance measures have also been used to harmonize other migration flow tables (Poulain (1999) and Poulain & Dal (2007)) in single time periods.

In this paper, the constrained optimization method is extended to alternative distance measures, constraint sets and generalized across a series of migration tables. First, appropriate data sources are selected using expert opinion. Sources that have reported counts that are considered to be insufficient or not available are ignored. Estimated correction factor to scale data from these sources would further enhance or depress existing unreliable patterns in reported values. Flows that were reported from reliable sending and receiving data sources are arranged into a set of migration tables (that may be non-square). Second, correction factors for at least one of the reliable data sources are set equal to one. The characteristics of this data source(s) will be used as the benchmark to scale all other reliable data. Third, estimate of all other correction factors for each selected data sources are determined using constrained optimization routines in statistical software. This is undertaken for 1) a series of selected migration flow tables over time and 2) a range of distance measures. Thus for each distance measure, a set of correction factors $(\mathbf{r}_t, \mathbf{s}_t)$ are estimated. The stability of these factors over time can be empirically summarized by considering $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^T = (\log(\mathbf{r}_t), (\log(\mathbf{s}_t))^T$. The variance within correction factors over time can thus be estimated,

$$\frac{\sum_{d=1}^{p} \sum_t (\boldsymbol{\theta}_{dt} - \bar{\boldsymbol{\theta}}_d)^2}{n-p}, \tag{4}$$

where n is the total number of correction factors over all time periods. Due to the asymmetry of scaling effects, the logarithmic transformation of correction factors are taken in the estimation of (4). This allows the variation between larger correction factors to have an equal effect as smaller correction factors. For the distance measure associated with the smallest variance, a new set of time constant factors $(\mathbf{r}, \mathbf{s})$ are estimated. This is undertaken by generalizing the distance function for an array of migration tables, m_{ijtk} with a dimension for time.

The final correction factors are applied to reported data as such,

$$y_{ijt} = \begin{cases} r_j m_{ijt1} & \text{if } r_j \text{ and } m_{ijt1} \text{ exist at time } t, \\ s_i m_{ijt2} & \text{if } s_i \text{ and } m_{ijt2} \text{ exist at time } t \text{ and } r_j \text{ does not,} \\ z_{ijt} & \text{otherwise,} \end{cases} \tag{5}$$

to create a series of migration flow tables y_{ijt} where z_{ijt} represents missing values. The application of correction factors in (5) is an alternative strategy to the approach suggested by Poulain (1993) who took an average of the scaled data. The correction of receiving data, when sending data are available, results in the distribution across a given column of migration flow table being preserved to that of the reliable reported data. This preference is undertaken for two reasons. First, receiving data is often believed to be of better quality (Erf (2007) and Raymer (2007)). Second, receiving data from some countries are highly regarded, and hence an alteration in their value might lead to implausible estimates. Scaled sending data is used when no reliable receiving data is available. This results in an altered distribution of flows across a row when compared with the original data. This alteration will be to greater effect than under an averaging of corrected flows, but provides estimates for counts in destinations where no reliable receiving data are available.

3.3 Impute Missing and Ignored Data

Spatial interaction models associated with Wilson (1970) have commonly been applied to mobility tables to expand the substantive understandings of studied transitions (refer to Fotheringham *et al.* (2000, p213-235) for a thorough discussion of the models). These traditionally employed mathematical algorithms to calibrate flow values to constrained origin and destination totals. Flowerdew & Aitkin (1982) and Willekens (1983) showed that a Poisson regression model with row and column dummy covariates are equivalent to constrained spatial interaction models for origin and destination totals,

$$\log \mu = \mathbf{X}\boldsymbol{\beta}, \tag{6}$$

where $Y \sim Po(\mu)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^O, \boldsymbol{\beta}^D \ldots)$ and $\boldsymbol{\beta}^O, \boldsymbol{\beta}^D$ are sets of origin and destination parameters respectively. Such models have been fitted to internal migration data using additional parameters for economic, geographical and population factors that may explain the size of migration flows, see for example Flowerdew & Lovett (1988) or Flowerdew (1991). These often lack a good fit as counts are aggregated over individual characteristics, such as age and sex, that are often useful in explaining people's movements (Congdon, 1991). This problem may be overcome by using a more flexible distribution assumption. Davies & Guy (1987) and Congdon (1989) suggested the use of a negative binomial distribution assumption to account for overdispersion effectively, $Y \sim NB\,(\log \mu, a)$, where the mean parameter $\log \mu$ is the same as in Equation (6) and a is the measure of dispersion.

In this paper, negative binomial models are applied to incomplete international migration flow tables. The dispersion parameter allows overdispersion in the observed harmonized data y_{ijt} generated by individual characteristics to be controlled for, and hence more realistic imputations for missing values, z_{ijt}. Covariates measured on aggregate levels are used to explain spatial interactions between countries. There are many theories that explain international migration, see for example Massey *et al.* (1993) or Greenwood & Hunt (2003). Data for economic, geographical and demographic factors suggested by these theories are often comparable across multiple nations and available from databases of international organizations.

Parameter estimates can be found by fitting spatial interaction models using the Expectation-Maximization (EM) algorithm of Dempster *et al.* (1977). This allows the observed likelihood to be augmented to account for the missing data, and thus when maximized, parameters are reflective of the complete data. As the algorithm is numerically stable, where the augmented likelihood increases at each iteration (Little & Rubin, 2002, p167), the asymptotic variance covariance matrix of parameters for the complete data can also be estimated. In this paper, estimates of this matrix are obtained using the Supplemented EM (SEM) algorithm of Meng & Rubin (1991). Refer to Little & Rubin (2002) or McCullagh & Nelder (1983) for a full discussion and details of the implementation of these algorithms. The SEM algorithm was written in S-Plus (available on request from the author) to provide parameter estimates, and their asymptotic variance covariance matrix, **V**, for negative binomial regression models. This required the `glm.nb` function in the MASS library (Venables & Ripley, 2002) for the M-step of the EM algorithm. Given the parameter estimates, the expected value for z_{ijt} can then be obtained. In addition, levels of the precision of these estimates can be derived through scaling covariate values in the model matrix by their estimated asymptotic standard errors and a Z-value based on 95% confidence level,

$$\log z_{ijt} \pm 1.96\mathbf{X}\mathbf{V}\mathbf{X}^{\mathbf{T}}. \tag{7}$$

These are incorporated with the harmonized flow values to provide comparable international migration data of all flows between selected countries.

4 Results

In this section the methodology is applied to real data in five parts. First, data for flows between the EU15 countries is outlined. Second, the count of migrants with unknown origins and destinations in the EU15 are presented. Third, data to and from selected countries is harmonized using a constrained optimization routine. Forth, covariates for a model to estimate missing and ignored data are outlined. Finally, estimates for parameters and their variance of a chosen model are calculated using the SEM algorithm.

4.1 EU15 Migration Data 2002-2006

International migration flow data may be obtained from a number of international orgainsations. One of the most comprehensive collections is provided by Eurostat (Kupiszewska & Nowok, 2008). Data are collected from individual national statistical institutes through a questionnaire on international migration statistics sent annually to 55 countries, organized by five organizations: Eurostat, United Nations Statistical Division, United Nations Economic Commission for Europe, Council of Europe (CoE) and International Labour Organization. Eurostat processes and disseminates data for the 37 European participants via their official database, (New Cronos) which is available online. The reported counts of these flows can also be found in publications of individual national statistical institutes, the CoE and SOPEMI reports of the Organization for Economic Co-operation and Development (OECD). Values of the same flows may not always be the same in all international organization databases. The cause of this difference is not known due to insufficient documentation (Kupiszewska & Nowok, 2008).

Data was obtained from the New Cronos web site (`http://epp.eurostat.ec.europa.eu`, accessed March 2008) for flows between EU15 nations in years 2002 to 2006. This set of countries was chosen due to the availability of literature on international migration statistics provided by national statistical institutes. In addition, a wide variety of the causes of incomparability in flow data are present.

Reported flow counts tended to be highest into and out of countries with the largest populations such as Germany, Great Britain, France, Italy and Spain. Flows between neighbouring countries, such as Netherlands and Belgium or Germany and Austria, tended to be higher than other values in the same row or column.

Of the 1050 cells (made from a 15×15 non-diagonal mobility table over 5 years), 225 cells had no reported values from either country. For 20 flows (out of a possible 210) there is no data reported in any year. In 870 cells, values from at least one reporting partner were present. In 332 cases data from both sending and receiving countries were available for which none reported the same value. As shown in Figure 1, for certain flows the difference in reported counts are constant over time. For other flows, variations in the differences between reported counts in and out of some nations such as Great Britain occur over time. In most cases the partner country reported fairly constant volumes of migrants, whilst British counts had more variation across time.

Some of the smallest difference occurred for flows between the Nordic nations of Sweden, Finland and Denmark. These countries all use registration systems to collect migration data for which a cooperation is in place, allowing migrants between them to be only registered in one country at a time (Herm, 2006). Consequently, data for the number of migrants sent from another Nordic nation is recorded by the country of destination, rather than origin and no measure of the amount sent is available. Small differences in the reported numbers are attributed to dual citizenship and time delays for migrations occurring at the end of the year (Nowok *et al.*, 2006).

4.2 Correction for Unknown Counts

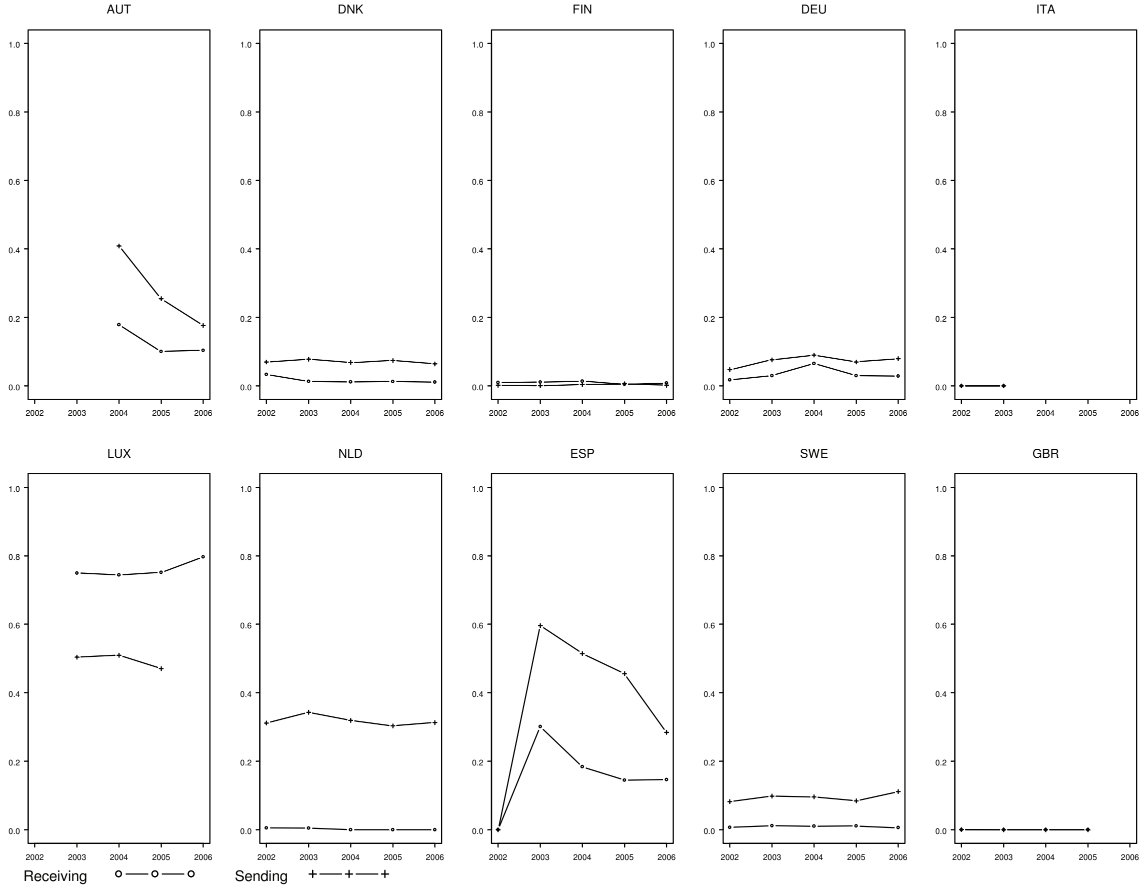

Figure 2: Proportion of Migrants Origin or Destinations Unknown for Available Receiving and Sending Data of EU15 in 2002-2006

Plots of the unknown counts as a proportion of total sending and receiving countries are shown in Figure 2 between 2002 and 2006. Totals for this calculation were given in the New Cronos database, which are themselves summations of all flows with both known and unknown origins and destinations. As with the flow data, counts are reported according to local definitions and data collection methods. For most countries, the plots demonstrate that sending data tended to have a greater proportion of unknown destinations in comparison with the unknown origins in receiving data, with the exception of Luxembourg. For some countries, such as Italy, Great Britain and Finland, the amount of unknown counts was small, or zero. Larger percentages are found for sending data of Luxembourg, Spain and the Netherlands. For Luxembourg, the large levels of unknown origin or destination are created by non-reporting of departures by leavers, and the non-collection of country of origin for people arriving, by the local municipalities from which national level data is aggregated (Perrin & Poulain, 2006a). For Spanish data, there is a notable change in the level of unknowns occurs between 2002 and 2003, increasing from 69 (and 6) migrants to 202,256 (and 38,339) received (and sent respectively). This pattern might be related to a switch in the data sources used to supply the data requested by the Joint Statistical Questionnaire on International Migration in 2001 (Breem & Thierry, 2006b). In the Netherlands, emigrants have to deregister from their municipal database when they leave the country with the intention to stay abroad for at least eight of the forthcoming twelve months. When people do not declare their departure, the register is later corrected without personal notification. For such administrative corrections, the country of destination is not known, creating the large unknown count (Erf *et al.*, 2006a).

All unknown counts are distributed to origins and destinations using the equations in (1). This reduces the difference between some reported counts, such as flows into Luxembourg where reported receiving data is almost always lower than sending data of corresponding origin countries.

4.3 Harmonization of Selected Data

Table 1: Erf (2007) Ratings of Migration Data for EU15 from 2002 to 2006

Country	Receiving			Sending		
	Timing	Completeness	Accuracy	Timing	Completeness	Accuracy
AUT	3	4	4	3	4	4
BEL	3	9	9	3	9	9
DNK	2(3)	4(4)	4(4)	3	4	4
FIN	2(4)	4(4)	4(4)	4	4	4
FRA	3	2	9			
DEU	2	4	4	2	4	4
GRC						
IRL	2	2	2	2	2	2
ITA	2(3)	3(3)	3(3)	4	3	3
LUX	2	3	3	2	3	3
NLD	3	4	4	4	4	4
PRT	4	9	9	3	2	2
ESP	2	3	3	2	3	3
SWE	4	4	4	4	4	4
GBR	4	2	2	4	2	2

0:Worst 1:Worse 2:Insufficient 3:Reasonable 4:Good 5:Excellent 9:Unknown.
Scores in parentheses are for non-national when national and non-nationals data is collected differently.
Countries labeled according to three-letter classification of ISO (2006).

Erf (2007) provided a subjective judgement of European migration flow statistics by

Table 2: Different Distance Metrics and Estimated Variance from 2002-2006 Data

Distance	$f(r_j, s_i \mid m_{ijk})$	Variance
Manhattan	$\sum_{i,j} \lvert r_j m_{ij1} - s_i m_{ij2} \rvert$	0.3217
Euclidean	$(\sum_{i,j} \lvert r_j m_{ij1} - s_i m_{ij2} \rvert^2)^{\frac{1}{2}}$	0.3846
Canberra	$\sum_{i,j} \frac{\lvert r_j m_{ij1} - s_i m_{ij2} \rvert}{r_j m_{ij1} + s_i m_{ij2}}$	0.2219
Clark	$\sum_{i,j} \frac{\lvert r_j m_{ij1} - s_i m_{ij2} \rvert^2}{(r_j m_{ij1} + s_i m_{ij2})^2}$	0.3434

the three characteristics: definitions of migration, measurement systems used and intended coverage. For the EU15 countries, ratings for both receiving and sending data between 2002 and 2006 are reproduced in Table 1. Ratings based on timing were judged by the degree of agreement with a twelve month timing criteria. This definition is recommended by the United Nations (UN) to reflect long term migrants who have changed their usual country of residence (UN, 1998). Ratings of completeness are based on the degree of under-registration suspected in the measurement systems. Scores for accuracy are based on the coverage of the target population and the collection and dissemination of data. Values for completeness and accuracy measurements were judged by considering the data sources used and experience with vital statistics. For most of the EU15 nations scores on completeness and accuracy of receiving and sending data were the same. Throughout the time period Greece failed to provide any receiving or sending data while France reported only receiving data. For three nations, Denmark, Finland and Italy, receiving data are collected differently for nationals and non-nationals, where the ratings for non-nationals are given in parentheses. All scores are constant over the 2002-2006 time periods.

Sub-tables of migration flows from data sources which were ranked with scores of at least reasonable for completeness and accuracy were created. As not all data from the reasonable providers was available throughout the time period, the dimension size of sub-tables and consequently the number of correction factors to be estimated changed in each year. Distance measures for flows between Nordic countries were ignored due to the data sharing agreement in place.

In each time period, correction factors were estimated to minimize a range of distance functions. This was undertaken using the `nlminb` routine in S-Plus 7.0. For data sources that scored good for timing, completeness and accuracy (according to Erf (2007)) the respective correction factors were constrained to one. This was done by setting the lower and upper bounds, required by the routine to 1.0. Bounds between 0.1 and 10 were imposed for all other correction factors. All initial parameter estimates for the function were set to 1.0. The range of distance functions ($f(r_j, s_i|m_{ijk})$) considered for the routine

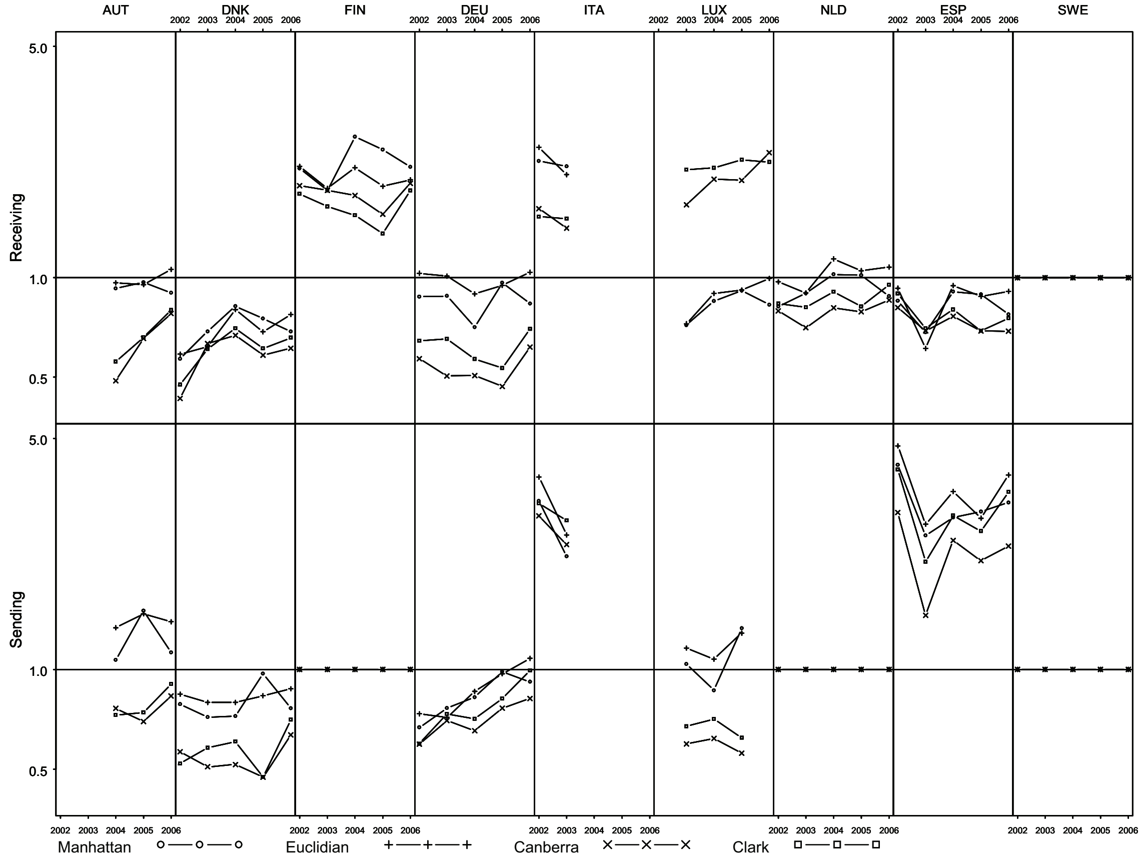

Figure 3: Receiving (r_j) and Sending (s_i) Correction Factors, 2002-2006 for Different Distance Functions

are shown Table 2. The first two measures were the Manhattan and Euclidean measures, (the latter equivalent to Equation (3). The third and fourth distance functions are based on the Canberra and Clark measures (Lance & Williams, 1967).

Estimates for the correction factors from these measures in each time period are given in Figure 3. For the first two measures, estimates tended to have similar values for each data source as they provided equal weighting for each double-counted cell in each migration flow table. The last two measures are also often very similar to each other and on occasions different to the previous two measures, as demonstrated by higher estimates in Luxembourg's receiving data correction factors. This was due to the weighting that both measures employ, allowing differences to be compared relative to the scaled reported data.

With a few exceptions, estimated correction factors tended to be similar over time and consistently greater or less than one. In a few cases, such as sending data from Luxembourg or Austria, the choice of distance measures would alter the direction of scaling. Spanish estimates fluctuated greatly in comparison with others, with values for most distance measures falling for 2003. This might be due to changes in the data collection methods from previous years.

Table 3: Estimated Correction Factors Over Series of Migration Tables

Country	r_j	s_i
AUT	0.6926	0.7594
DNK	0.6357	0.5751
FIN	1.8096	1.0000
DEU	0.5637	0.7067
ITA	1.6502	2.8339
LUX	1.9691	0.6665
NLD	0.8227	1.0000
ESP	0.7715	2.6730
SWE	1.0000	1.0000

Comparison across all distance measures for the selected data sources from the EU15 are shown in Table 2. The smallest variation over time in the logarithm of correction factors, calculated using Equation (4) is that of the Canberra measure. As definitions and collection methods of all the reported data used in the estimation are assumed to be unchanged, the measure that possessed the smallest variation was regarded as the most reasonable for a constrained optimization for the given data. Thus a set of time constant correction factors for each data source, (over the entire series of tables) was estimated from an constrained optimization on an generalized Canberra distance measure,

$$f(r_j, s_i|m_{ijtk}) = \sum_{i,j,t} \frac{|r_j m_{ijt1} - s_i m_{ijt2}|}{r_j m_{ijt1} + s_i m_{ijt2}} \tag{8}$$

This optimization was undertaken with constraints on correction factors with timing criteria rated as good by Erf (2007). Estimates of correction factors are given in Table 3. These were applied to the criteria of (5) to obtain harmonized values for flows to and from all reliable data sources.

4.4 Covariates for Model Based Imputations

A negative binomial regression model was fitted by implementing the EM algorithm to harmonized international migration flow data between 2002 and 2006. After harmonized flow values were obtained for data sources that were considered reliable, 819 cells (of 1050) had observed counts. In 30 (of the 210) flows there were no observations of harmonized data in any years. This was greater than the reported data (20), as some values are ignored due to their poor quality.

In order to provide reasonable imputations, data on nine factors were collected to reflect differing economic determinants, geographical characteristics and populations between origins and destinations for international migrants. Where possible, information across time was taken to help reflect trends in migration flow counts.

Four covariates on economic systems were constructed: the origin-destination ratio of Gross National Income (GNI) per captia and Gross Domestic Product (GDP), the logarithm of the total value of trade for each corresponding flow and a dummy variable for the circulation of the Euro currency in both origin and destination countries.

Data for GNI and GDP were obtained from the World Bank, World Development Indicators Database (`http://www.worldbank.org/data`). Values that used a purchasing power parity adjustment to account for differences in relative living costs and inflation were used. A per capita measure was taken for GNI to reflect a macro measurement of differences in wages. GDP was measured on a national level to reflect differences in economies income and output. The logarithm of this ratio was taken due to the higher level of asymmetry created by the comparison of large economies such as Germany, France and Great Britain to smaller nations such as Luxembourg. A covariate measure on trade was collected in order to reflect economic linkages between nations. Data for the value of all commodities imported into each country for all origin nations was obtained from the UN Commodity Trade Statistics Database (`http://comtrade.un.org/`). A final economic covariate measure was constructed to represent countries using the Euro, to potentially explain higher flows between countries where levels of economic and political integration might be even greater than flows from other EU15 nations due to a common currency.

Two measurements of geographical links were created: distance and contiguity. A weighted distance between two countries was obtained from Mayer & Zignago (2006). Measurements are calculated in kilometres between the principal cities of countries weighted by their population size and thus account for the uneven spread of population across a country. A separate dichotomous measure for contiguity was taken as internal migration studies have sometimes shown its impact to be distinct from that of distance (Flowerdew & Lovett, 1988). Data for this variable was obtained from Stinnett *et al.* (2002) where countries separated by land, river border or 12 or less miles of water are considered contiguous.

Three covariates on population were considered: size, migrant stocks and language. Comparisons between multiple nations used the sum of origin and destination populations. This manipulation was used to order to control for higher migration flows between countries with large populations such as Germany and France. Population data was also obtained from the World Bank, World Development Indicators Database. An origin-destination migration stock table was derived from Parsons *et al.* (2005) who complied a global bilateral database from the 2000 round of population censuses. Covariates on languages were considered to further reflect social and linguistic similarities. These where derived from a European Commission's Eurobarometer survey on European's and their Language (`http://ec.europa.eu/public_opinion`). Variables for the official languages used in more than one of the EU15 (English, French and German) were based on the surveys estimates of the knowledge of each tongue as a foreign language in each nation. The product of origin and destination language prevalence were then calculated, after setting values for foreign languages levels in countries, where it was officially spoken, to 100 percent (lower levels were recorded as a non-native speaking survey respondent considered the official language as a foreign tongue). For example, values representing the commonality of English and French for the Netherlands to Great Britain flow were 0.8700 and 0.0667 respectively, indicating a higher overall level of English in the two nations. An additional continuous covariate for time was also added to account for changes in the level of migration flows during the time period, and the correlation amongst repeated counts of the same origin-destination combination over time.

4.5 Complete Migration Flow Tables

The `stepAIC` function in the MASS library (Venables & Ripley, 2002) of S-Plus 7.0 was used to select covariates based on the observed (harmonized) data. The function operated by examining the inclusion of potential covariates by their contribution to the Akaike Information Criterion (AIC) (Akaike, 1973) of the model by performing a stepwise search in both directions i.e., adding and dropping variables in the model. Included as a precondition for the scope of models to be searched were the origin and destination covariates. Covariates for distance, contiguity and German language were found to be ineffective in reducing the AIC. The selected model was then re-fitted using the SEM algorithm to provide parameter estimates and asymptotic variance-covariance matrix that account for the incomplete data.

Convergence of parameter estimates, from the EM part of the algorithm, was obtained after 33 iterations with a tolerance level of 10^{-6}. The asymptotic variance-covariance matrix took only six iterations to converge with a stopping criteria of 10^{-3}. More stringent stopping criteria were attempted but resulted in non-convergence for some elements of the rate of change matrix estimated in the SEM algorithm. This problem was suspected to be caused by the methods used to estimate parameters in the M-step of the EM algorithm. As the negative binomial distribution does not belong to the exponential family, the dispersion parameter was estimated using asymptotic approximations based on linearizations from a Newton-Raphson routine in `glm.nb`. Such fitting methods may create numerical inaccuracies in comparison to alternative methods such as Iteratively Rewighted Least Squares used for estimating the other parameters in the model (Meng & Rubin, 1991). As the rate of change matrix consisted of 1296 elements from 37 parameters (one constant, fourteen origins, fourteen destinations, seven other economic and population measures and the dispersion) numerical inaccuracies were likely with higher tolerance levels.

The variance-covariance matrix estimated using the rate of change matrix from the converged SEM was symmetric when rounded to two decimal places. As Meng & Rubin

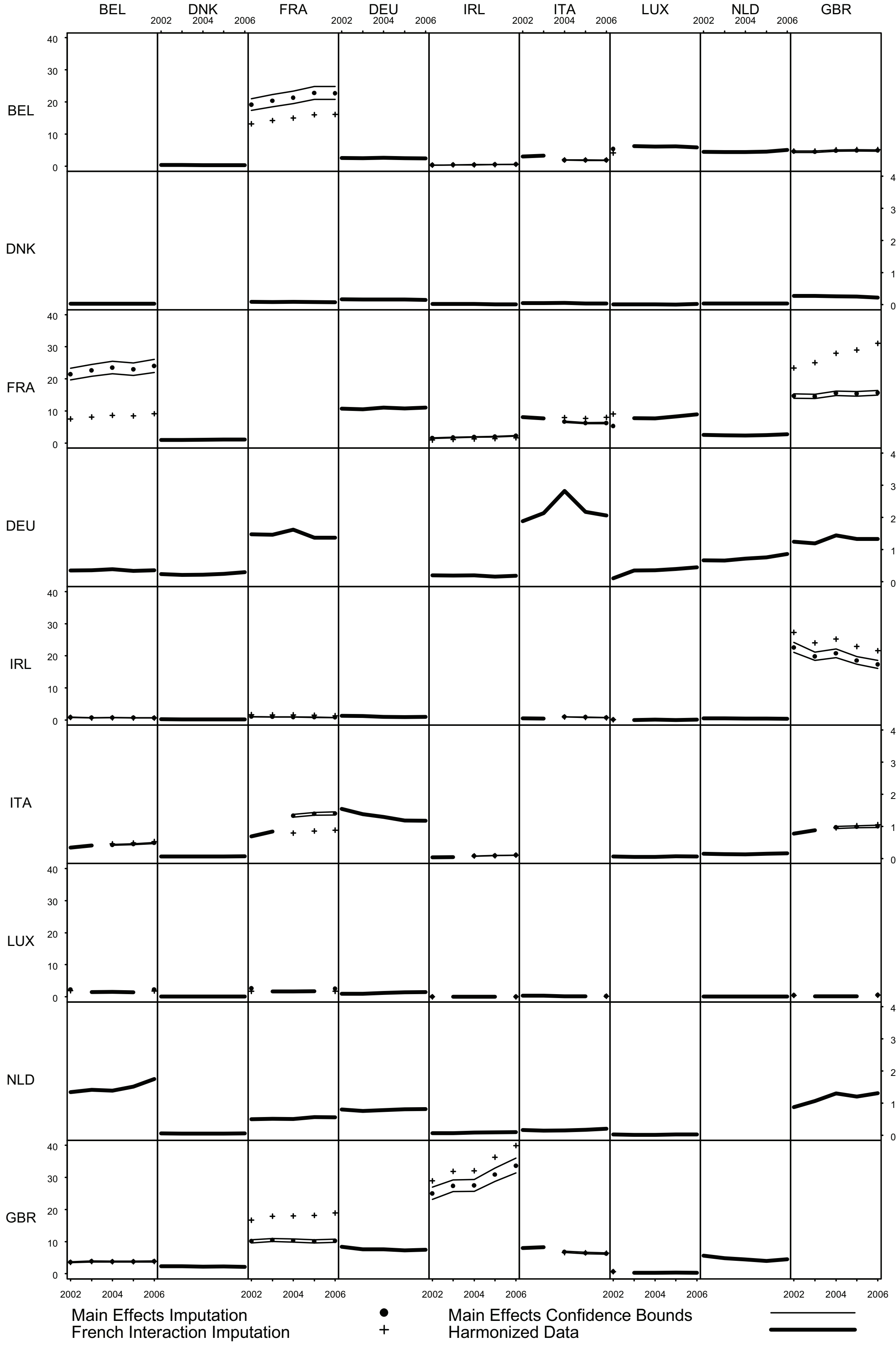

Figure 4: Imputations and 95% Confidence Bounds of Estimated Migration Flows (000's) between EC9 nations, 2002-2006.

Table 4: Covariate Estimates of Spatial Interaction from Selected Model

Covariate	Type	Time Varying	$\exp(\boldsymbol{\beta})$	$\mathrm{Var}(\boldsymbol{\beta})$
GNI	Ratio	Yes	6.7608	0.1281
GDP	Logarithm of Ratio	Yes	2.3277	0.1869
Trade	Logarithm	Yes	1.3444	0.0015
Euro	Dichotomous	No	1.4340	0.0118
Stock	Logarithm	No	1.8185	0.0006
French	Percentage	No	3.3335	0.0615
English	Percentage	No	0.3996	0.1368

(1991) noted, this is an important check for computational errors. The parameter estimates and the variance-covariance matrix were used to derive fitted values and their 95% confidence bounds using Equation (7). These are shown by the circles and thin lines in Figure 4 for a sub-table of the EU15, consisting of the nine countries of the European Communities (EC9) before 1981. Fits to observed data, (not shown on Figure 4) all appeared reasonable. Larger flows, such as from Germany and Great Britain to Spain, were lower than the harmonized counts. This may be due to other factors, such as retirement migration flows, lacking representation in the selected model.

The exponentiated parameters and their variances, found by the SEM algorithm, are given in Table 4 for the seven covariates selected by the stepwise model selection procedure. In addition to determining estimates of missing flows, their values gave some element of a substantive understanding for spatial interaction within the EU15.

The 14 exponentiated origin parameter values (not shown in Table 4) provided a measure of the level of attraction of migration flows, in comparison to Austria which was used as a reference category. Values varied from 18.7665 and 8.6941 for Germany and Great Britain to 0.5039 and 0.6915 for Luxembourg and Ireland respectively (Austria's parameter was set to one). Exponentiated destination parameter values (where Austria was again the reference category) varied from 11.2861 and 8.0532 for Germany and Great Britain to 0.8178 and 0.9950 for Ireland and Denmark respectively.

The estimated exponentiated parameters effects for economic factors (6.7608 for the ratio of GNI per capita, 2.3277 for the logarithm of the ratio of GDP, 1.3444 for the logarithm of trade volume and 1.4340 for Euro region), logarithm of migrant stocks (1.8185) and French prevalence (3.3335) were all greater than unity implying higher levels of these covariates were associated with higher migration flows, conditional upon the value of all other covariates. Exponentiated coefficients estimates for English prevalence (0.3996) was less than unity indicating higher levels in their covariates were associated with lower migration flows, given all other variables are controlled for. This might be due to low covariate values being determined between countries with high migration flows. For example, the value of English prevalence for a migrant moving from Sweden to Great Britain was 0.8900, compared to 0.5607 for (more popular) moves from Sweden to Finland. Similar problems did not occur with other languages, which tended to have much smaller levels of commonality throughout most countries. The dispersion parameter (a) was estimated to be 8.2560 with standard error 0.3640. A Z-test provided strong evidence that $a > 0$, suggesting a negative binomial model was more appropriate than a equivalent Poisson model.

Analysis of the fits from the main effects model in Figure 4 showed reasonable imputations for most of the previously missing cells. An exception was the selected flows to and from France. For example, the number of migrants sent from France to Belgium was higher than movements to other neighbouring countries of greater population size and economic power, such as Spain or Germany. For these countries, fitted values to and from France tended to be greater than the harmonized values, creating large residuals. This might be caused by the general nature of the main effects model for the whole EU15 region. However, some factors included in the model may vary substantially for migration flows to or from individual nations.

To this end, the `stepAIC` function was run again with an extended scope of models to consider all two-way interaction except the origin-destination interaction. This was not included as for some levels, such as the flows between Britain and France, no data existed and hence such a parameter could not be identified. The stepwise function selected two new main effects (German and distance) and 24 new interaction covariates. From the total of 26 new covariates many involved origin or destination interactions and hence multiple levels, producing a total of 243 new parameters. Consequently many of these parameters were unidentified and imputations were unreasonable.

Whilst a model with many interactions and multiple parameters may not be plausible for a migration table involving many countries, interactions for single countries could be constructed to effectively improve model imputations where deemed necessary. The second set of imputations in Figure 4 were obtained by creating interaction variables for the 11 parameters (including three languages) with France as both an origin and destination. The 22 additional covariates where considered by the stepwise model fitting algorithm. The AIC of final selected interaction model was 15,280, a reduction in comparison to the main effects model (15,836) but with more parameter estimates (from 37 to 46). Of these, six were new interaction covariates and three were new main effects (for population, distance and time). The additional main effect covariates may have been included as higher level interactions with other covariates or with France (as an origin or destination) were effective and hence its main effects were also useful in explaining the spatial interactions. Alternatively, parameter estimates from the original main effects model were altered by the inclusion of interactions and thus more main effects were added to cover the change in model fit. Of the six new interactions, five (GNI ratio, population sum, the Euro, stock and distance) were with France as an origin and one (stock) were with France as a destination. Their inclusion indicated evidence that factors had different effect for flows to or from France in comparisons to their main effects for the EU15 region.

All parameters were identifiable and led to a noticeable improvement in the imputations of cell values in the French row and column of Figure 4 respectively. This is best demonstrated for flows from Italy to France, where imputations in later years follow neatly from harmonized data in the first two time periods. In addition, flows from Belgium, which where considered unusually high fell, whilst flows to and from larger countries such as Great Britain increased.

5 Summarizing Remarks and Discussion

The estimation of international migration is a complex process. The multidimensional nature of migration and the differences in the forms of measurement and data collection, make any estimation attempt difficult. In this paper, a focus on all international migration flows between a set of multiple countries was taken. To obtain these estimates, problems in inconsistent reported flow counts from reliable data sources were first addressed using a constrained optimization procedure. Estimates of missing data and measure of their precision were obtained by fitting a negative binomial model using the SEM algorithm. The resulting estimates are considered comparable across all flows.

Data from different countries and over different time periods can be easily incorporated to the methodology illustrated in this paper. The non-linear optimization routines used for the harmonization process can be altered to incorporate changes in constraints, the use of alternative distance measures, estimates for extra parameters if data production techniques change and more realistic bounds for correction factors (that might be supplied by data experts) to be set. Routines might also be easily constrained to harmonize data to an alternative timing criterion if available in the data source(s) of the studied set of countries.

Models used by the SEM algorithm to impute estimates and their precision can be altered depending on the users needs. As demonstrated for France, experts can help inform the model building process. Imputations that may require further parameters in comparison to a model for the complete migration system can be added where deemed necessary. Further main effects and redefining the origin-destination relationships in existing covariates might also be further explored. For example, comparative measures of unemployment or climate could be utilized if reliable data are available. Information on population groups, such as students, may also be beneficial to model fits. Its inclusion might be interacted with a dummy covariate to indicate if the population group has or has not been included in the reported data.

Despite the common occurrence of missing data in international population mobility tables, the application of the EM algorithm is sparse. The EM algorithm allows wide range of techniques for the statistical modelling of mobility tables to be applied. In doing so, models are able to account for missing data and impute missing cell values based on statistical assumptions and covariate information drawn from migration theory. In addition, measures of precision for imputations can be derived using the SEM algorithm. Previous methods for imputing data in international tables have tended to focus on mathematical relationships of different data sets rather then statistical solutions. Parsons *et al.* (2005) used an entropy measure between different migrant stock definitions, whilst Poulain (1999) scaled other data sources in place of missing flows. More statistical approaches of Raymer (2007), and extensions of this work (Raymer (2008), Raymer & Abel (2008) and Brierley *et al.* (2008)) estimate missing model components, rather than the flows directly. These are reliant on marginal totals being known, or easily estimated. However, as with individual flows, comparable reported values for these totals are difficult to obtain due to differences in data collection methods and definitions.

In this paper, as a prelude to the estimation of correction factor, counts of known migrants with unknown origins or destinations were accounted for by distributing these flows according to the existing flow information. The allocation of unknown counts assumes that information on migrant's origins or destinations are missing at random. If certain types of migration, such as inter-continental moves, are more likely to be reported then this allocation would discriminate against more local moves. If available, expert opinion could moderate this distribution by weighting the numerators in (1) appropriately.

The harmonization of flow data assumed the differences between reported data from reliable sources were non-random. This assumption could be modified by considering the estimation of flows in the Bayesian paradigm, where reported data might be considered as observations from an underlying negative binomial distribution with a mean parameter for receiving and sending migration tables ($\log \mu$ in Equation (6)) scaled by r_j and s_i respectively. Prior distribution on the correction factors would hence allow variation in the differences of data production techniques between different countries. This variation would also be fully reflected in the posterior distribution for missing cells.

The negative binomial regression model proved to be an effective tool to deal with overdispersion of the data. The use of alternative error assumptions such as a Poisson distribution would have lead to worse fitting models and non robust standard errors. The building of models relied upon comparisons of their AIC calculated from the observed data, rather than the complete data. As Cavanaugh & Shumway (1998) noted it is more desirable to fit a model based on the complete data for which models are originally postulated for and hence include information on the missing data. Criteria, such as the AIC-cd of Cavanaugh & Shumway (1998) and KIC-cd of Seghouane *et al.* (2005), allow the calculation of the separation between the fitted model for the complete data and the true or generating model. Both criteria require models to be fitted using the SEM for potential models, and hence to find a suitable model would require a greater computational time than the stepwise model selection routine.

Comparable international migration flow data are needed by researchers working on identifying, understanding and monitoring migration flows. Governments and planners can also use more comparable estimates to help forecast the demand for services that are created by population changes, for which the role of international migration can have a significant influence. The methodology outlined in this paper provides a relatively flexible technique to overcome the problems of inconsistencies and incompleteness in international migration data.

6 Acknowledgements

This work was undertaken with financial support of the Economic and Social Research Council. I am very grateful to Corrado Giulietti, Dr. James Raymer and Professor Peter W.F. Smith and for their comments and suggestions.

References

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. *Pages 267–281 of: International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR.*

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., & Hugo, G. 2002. Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **165**(3), 435–464.

Breem, Y., & Thierry, X. 2006a. Counrty Report: France. *Pages 457–466 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM).* Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Breem, Y., & Thierry, X. 2006b. Counrty Report: Spain. *Pages 447–445 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM).* Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Brierley, M.J., Forster, J.J., McDonald, J.W., & Smith, P.W.S. 2008. Bayesian Estimation of Migration Flows. *Chap. 7, pages 149–174 of:* Raymer, J., & Willekens, F. (eds), *International Migration in Europe.* Chichester, United Kingdom: Wiley.

Cavanaugh, J.E., & Shumway, R.H. 1998. An Akaike Information Criterion For Model Selection In The Presence Of Incomplete Data. *Journal of Statistical Planning and Inference*, **67**(1), 45–65.

Congdon, P. 1989. Modelling Migration Flows between Areas: An Analysis for London Using the Census and OPCS Longitudinal Study. *Regional Studies*, **23**(2), 87–103.

Congdon, P. 1991. General Linear Modelling: Migration In London And South East England. *Chap. 7, pages 113–136 of:* Stillwell, J., & Congdon, P. (eds), *Migration Models: Macro and Micro Approaches*. London, England: Belhaven Press.

Davies, R.B., & Guy, C.M. 1987. The Statistical Modeling Of Flow Data When The Poisson Assumption Is Violated. *Geographical Analysis*, **19**(4), 300–314.

Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Erf, R. van der. 2007 (June). *Feasibility Study and Associated Work Plan.* Deliverable 1.2. Netherlands Interdisciplinary Demographic Institute (NIDI), The Hague, Netherlands.

Erf, R. van der, Heering, L., & Spaan, E. 2006a. Counrty Report: Netherlands. *Pages 553–564 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Erf, R. van der, Heering, L., & Spaan, E. 2006b. Statistics on Asylum Applications. *Chap. 10, pages 249–319 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Flowerdew, R. 1991. Poisson Regression Models Of Migration. *Chap. 6, pages 92–113 of:* Stillwell, J., & Congdon, P. (eds), *Migration Models: Macro and Micro Approaches*. London, England: Belhaven Press.

Flowerdew, R., & Aitkin, M. 1982. A Method Of Fitting The Gravity Model Based On The Poisson Distribution. *Journal of Regional Science*, **22**(2), 191–202.

Flowerdew, R., & Lovett, A. 1988. Fitting Constrained Poisson Regression Models To Interurban Migration Flows. *Geographical Analysis*, **20**(4), 297–307.

Fotheringham, Stewart A., Brunsdon, Chris, & Charlton, Martin. 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage Publications Ltd.

Goldstein, S. 1976. Facets of redistribution: research challenges and opportunities. *Demography*, 423–434.

Greenwood, M.J., & Hunt, G.L. 2003. The Early History Of Migration Research. *International Regional Science Review*, **26**(1), 3.

Herm, A. 2006. Recomendations on International Migration Statistics and Development of Data Collection at an Individual Level. *Chap. 2, pages 77–107 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

ISO. 2006. *International Standard ISO 3166-1, Codes for the representation of names of countries and their subdivisions.* International Organization on Standardization.

Kelly, J.J. 1987. Improving the Comparability of International Migration Statistics: Contributions by the Conference of European Statisticians from 1971 to Date. *International Migration Review*, **21**(4), 1017–1037.

Kraler, A., Jandl, M., & Hofmann, M. 2006. The Evolution of EU Migration Policy and Implications for Data Collection. *Chap. 1, pages 35–75 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM).* Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Kupiszewska, D., & Nowok, B. 2008. Comparability of Statistics On International Migration flows In The European Union. *Chap. 3, pages 41–73 of:* J. Raymer, F. Wiilekens (ed), *International Migration in Europe: Data, Models and Estimates.* London, England: Wiley.

Lance, G.N., & Williams, W.T. 1967. Mixed-Data Classificatory Programs I - Agglomerative Systems. *Australian Computer Journal*, **1**(1), 15–20.

Little, R.J.A., & Rubin, D.B. 2002. *Statistical Analysis With Missing Data.* Wiley.

Massey, D.S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J.E. 1993. Theories of International Migration: A Review and Appraisal. *Population and Development Review*, **19**(3), 431–466.

Mayer, T., & Zignago, S. 2006. Notes on CEPIIs distances measures. *Centre dEtudes Prospectives et dInformations Internationales (CEPII), Paris.*

McCullagh, P., & Nelder, JA. 1983. *Generalized Linear Models.* London, England: Chapman Hall.

Meng, X.L., & Rubin, D.B. 1991. Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm. *Journal of the American Statistical Association*, **86**(416), 899–909.

Nowok, B., Kupiszewska, D., & Poulain, M. 2006. Statistics on International Migration Flows. *Chap. 8, pages 203–233 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM).* Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Parsons, C. R., Skeldon, R., Walmsley, T. L., & Winters, L. A. 2005. Quantifying the International Bilateral Movements of Migrants. *8th Annual Conference on Global Economic Analysis, Lübeck, Germany, June*, 9–11.

Perrin, N. 2006. Counrty Report: Ireland. *Pages 467–489 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM).* Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Perrin, N., & Poulain, M. 2006a. Counrty Report: Luxembourg. *Pages 519–527 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on*

International Migration (THESIM). Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Perrin, N., & Poulain, M. 2006b. Counrty Report: United Kingdom. *Pages 645–655 of:* Poulain, M., Perrin, N., & Singleton, A. (eds), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Poulain, M. 1993. Confrontation des Statistiques de migrations intra-européennes: Vers plus d'harmonisation? *European Journal of Population/Revue européenne de Démographie*, **9**(4), 353–381.

Poulain, M. 1999 (May). International Migration within Europe: Towards More Complete and Reliable Data. Conference of European Statiticsans. Joint Economic Commission for Europe (ECE) and Eurostat, Perugia, Italy.

Poulain, M., & Dal, L. 2007 (March). *Estimation of All Flows Within the Intra-EU Migration Matrix*. Deliverable. GDAP-UCL, Louvain-La-Neuve, Belguim.

Raymer, J. 2007. The Estimation of International Migration flows: A General Technique Focused on the Origin–Destination Association Structure. *Environment and Planning A*, **39**(4), 985–995.

Raymer, J. 2008. Obtaining an overall picture of population movement in the European Union. *Chap. 10, pages 209–234 of:* Raymer, J., & Willekens, F. (eds), *International Migration in Europe*. Chichester, United Kingdom: Wiley.

Raymer, J., & Abel, G.J. 2008 (March). The MIMOSA Model for Estimating International Migration Flows in the European Union. Joint UNECE/Eurostat Work Session on Migration Statistics. United Nations Statistical Commission And European Commission Economic Commission For Europe Statistical Office Of The European Communities (EUROSTAT), Unite.

Salt, J. 1993. *Migration and population change in Europe*. Tech. rept. 19UNIDIR/93/23. United Nations Institute for Disarmament Research, (UNIDIR), New York, New York.

Seghouane, A.K., Bekara, M., & Fleury, G. 2005. A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal Processing*, **85**(7), 1405–1417.

Stinnett, D.M., Tir, J., Schafer, P., Diehl, P.F., & Gochman, C. 2002. The Correlates of War Project Direct Contiguity Data, Version 3. *Conflict Management and Peace Science*, **19**(2), 58–66.

UN. 1998 (September). *Recommendations on Statistics of International Migration. Revision 1*. UN.

Venables, W.N., & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer New York.

Willekens, F. 1983. Log-Linear Modelling Of Spatial Interaction. *Papers in Regional Science*, **52**(1), 187–205.

WILLEKENS, F. 1994. Monitoring international migration flows in Europe. *European Journal of Population/Revue européenne de Démographie*, **10**(1), 1–42.

WILSON, A.G. 1970. *Entropy In Urban And Regional Modelling*. Pion, London.