University of Southampton

Faculty of Law, Arts & Social Sciences

School of Social Sciences

International Migration Flow Table Estimation

by

Guy J. Abel

Thesis for the degree of Doctor of Philosophy

April 2009

To Ed and Diana

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) type 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.

Contents

List of Figures

List of Tables

Declaration Of Authorship

I, Guy Jonathan Abel, declare that the thesis entitled International Migration Flow Table Estimation and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: ....

April 2009

Acknowledgements

This work was undertaken with financial support of the Economic and Social Research Council (PTA No-031-2004-00274).

This thesis would not have been possible without the support of many people. I would like to express my sincere gratitude to:

- My supervisors, James Raymer and Peter Smith, who were abundantly helpful and offered invaluable advice, assistance and support throughout my studies in Southampton.
- Members of the Division of Social Statistics who taught me during my Masters year and encouraged me throughout the doctorate.
- Corrado Giulietti and Alessandro Mennuni for their help with the more mathematical parts of this thesis, reminding me of areas from my undergraduate degree that I had forgotten and showing me the ropes with Matlab.
- Tom King and Corrado Giulietti (again) who read previous drafts of this thesis, gave some great advice and helped me pull things together.
- The research support staff in iSolutions (and Corrado Giulietti) who helped me with getting to grips with the iridis computer cluster.
- The support staff in the Division of Social Statistics and S3RI who helped me with the administrative side of things.
- John Aston for inviting me to visit him at the Institute of Statical Sciences at Academia Sinica in Taipei. The visit was an amazing experience and was also the place I developed a lot of the ideas for this thesis.
- My parents, Ed and Diana Abel, who read earlier drafts and have given me their unconditional support and belief throughout my education.
- My brothers and sister, and their partners, who continuously provided support, places to sleep and paid for things I couldn't afford.
- Claire Cheng 鄭又慈 for all the 加油！
- My fellow research students and house mates (and the cross-section of the two) for all their friendship, support and good times.

Chapter 1

Introduction

Migration flow data inform policy makers, 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 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 help concerned parties to manage policy and understand people's movements better. 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. This thesis develops steps towards these ends, introducing a methodology for the estimation of international migration flow tables of comparable data.

This introductory chapter commences with an overview of international migration flow data. The lack of comparability in flow data can be grouped into two areas: inconsistencies and incompleteness. These problems have lead to the development of estimation methods for the provision of comparable data by previous researchers. The next section discusses migration flow tables. Analysis of international migration tables of comparable data have a number of discussed advantages which motivate their study throughout this thesis. The succeeding section describes the aims and scope of this thesis. Included are a set of desirable criteria for methodologies to estimate international migration flow tables of comparable data. These criteria are used to evaluate estimation techniques (including the one developed in this thesis) and help determine the comparability of resulting estimates. Finally, a summary of the thesis structure is given.

1.1 International Migration Data

Migration can be measured as either a flow or stock. Data for migration flows quantify the magnitude of population movements between selected countries during a specified time period (usually one year). Migrant stock data quantify the size of immigrant populations. This thesis concentrates on the first of these measures.

International migration flow data often lack adequate measurements of volumes, direction and completeness between nations (Kelly, 1987; Salt, 1993; Willekens, 1994; Nowok et al., 2006). The lack of comparability in flow data can be traced to a number of causes. First, migration is a multi-dimensional process (Goldstein, 1976) involving a transition between two states. Consequently, movements can be reported by sending or receiving countries. When data collection methods or measurements used in these countries differ, the reported counts do not match. Second, international migration flow data are typically collected by individual national statistics institutes in each country. Institutes have developed measures of migration solely suitable to their domestic priorities. These are often produced within a legal framework, and hence alterations to their collection are difficult to implement. Third, in many countries, migration data collection systems do not exist. In other countries, collection methods (such as passenger surveys) may provide inadequate means to report flows at the levels of detail required by some data users. Finally, the nature of international migration continues to change. In recent decades movements have become more global, occurring at faster rates and diversifying into a greater range of migration types, such as migration for short periods of time, for retirement or for political asylum (Castles and Miller, 2003, p7-9). National statistics institutes are often unable to adapt data collection and measurement procedures to provide users with information on such changes.

Difficulties in producing international migration flow statistics creates multiple problems in obtaining comparable data needed for a better understanding of population change and behaviour. These problems can be grouped into two areas: inconsistencies and incompleteness. Inconsistencies in reported values, for the same flow, occur due to different measurements and data collection systems. Incompleteness in reported values occurs when national statistics institutes do not collect or disseminate data. Estimation techniques, such as those applied in this thesis, can be used to overcome these problems. These techniques often require knowledge of the collection methods and measurement in each data source, assumptions regarding the difference between reported counts and statistical models for the imputation of missing flows.

1.2 International Migration Flow Tables

Data on migration between a set of regions are commonly presented in a square table with off diagonal entries containing the number of people moving from any given origin to any given destination. These are known as migration tables or matrices. The analyses of tables of comparable international migration data have a number of advantages. First, they allow a fuller understanding of population behaviour and change in comparison to other migration measures. For example, the study of a net migration measure cannot differentiate reported counts by migrant origins or destinations (Rogers, 1990). Second, international migration tables provide details on the propensity of movements across multiple countries. Consequently, the contributions made by each nation to a system of migration can be easily identified. An alternative analysis of migration, such as flows into a single country or net migration cannot account for this heterogeneity. For example, when modelling the movements into a single country over time, a similar country may undergo a period of immense growth, drawing migrants away from the country of study. The analyses of flows for the country of study may be able to explain fewer migrants sent from this high growth country but may fail to account for a fall in its relative attraction to potential migrants from other origins. Third, the analysis of migration tables allows the possibility for counts to be divided into sub-tables based on individual characteristics of migrants such as age and sex. These additional dimensions, as with origin and destination, allow the analysis of migration flow tables containing information in a whole system of movements, furthering the possibility for insights that may have been confounded by more conventional methods of analysis. Finally, migration tables may be considered as part of a wider account of demographic data. Rees (1980) noted that national account statistics of financial stocks and flows have served economists well in their modelling activities, encouraging users to compare data for consistencies, check for inadequacies and force analysts to attempt to match available data with a conceptual model. A demographic account of population stocks and flows would lead to similar improvements.

1.3 Thesis Aims and Scope

The study of transition patterns, such as migration flows, generally involves three steps (Rogers, 1980). First, data are collected and missing observations estimated. Second, appropriate rates and probabilities are calculated. Finally, simple projections of the future conditions that would arise were probabilities to remain unchanged are generated. This thesis concentrates on the first of the three stages, with the aim of providing a methodology for the estimation of comparable data for international migration flow tables.

The methodology developed in this thesis addresses the two fundamental data problems of inconsistencies and incompleteness. In order to make observed data consistent, a constrained optimization procedure is used. Such procedures have been applied to harmonize international migration data in previously proposed methodologies, such as Poulain (1993). These are reviewed in Chapter 3, and extended using a variety of distance functions and constraints in Chapter 4. There exists a range of model based methods to deal with missing data in statistical literature, see for example the seminal text of Little and Rubin (2002). One such method, the Expectation-Maximization (EM) algorithm is applied to harmonized international migration flows in Chapter 5. This method is chosen because of its effectiveness and relative simplicity in application compared with other missing data techniques. This new application of a popular missing data technique allows the imputations for missing migration flows to be estimated. In Chapter 6, the asymptotic variance-covariance matrix for parameter estimates is obtained using the Supplemented EM (SEM) algorithm of Meng and Rubin (1991). When this extended algorithm is applied to migration flow data, measures of precisions for imputations from Chapter 5 can be derived.

The methodology developed in this thesis, and alternative frameworks presented in Chapter 3, will be evaluated with respect to eight desirable criteria for methodologies in estimating international migration flow tables of comparable data. These are shown in Table 1.1. Criteria are divided into two groups: properties of estimates and properties of the methodology for their estimation. These are not mutually exclusive or exhaustive, and data users may require further criteria in estimates or the methodology. However, they do provide guidance for comparisons between estimates and their frameworks which may otherwise be difficult to evaluate.

Table 1.1: Desirable Criteria for Methodologies in Estimating International Migration Flow Tables of Comparable Data.

Estimates:
Complete
Consistent
Reliable
With associated precision measures
Methodology:
Model based imputations for missing data
Incorporate expert opinion
Easily replicable
Flexible to different time periods and regions

The first three criteria for estimates were originally outlined by Willekens (1994) for future work on combining data sources to create a statistical data base of international migration flows. When present, these characteristics will result in comparable data which can allow users to better understand population behaviour and change. One additional desirable criterion, for a measure of precision in the estimates, is also given. As estimation techniques are used to provide comparable data, an associated measure of precision can further aid data users to understand the possible variation associated with estimated values.

Willekens (1994) also suggested that a methodology for estimating a data base of comparable international migration flows should allow for the imputations of missing values to be based on models of migration and incorporate expert opinion. Models of migration counts create a description of each flow in the table in relation to other data. Once the model is specified, it may be used to impute or update data while preserving imposed structures or constraints. In addition, model based imputation methods can allow estimates to be based on likelihood methods, to obtain the most likely estimates given the data. Estimation methodologies can be aided by the inclusion of expert opinion, which may provide a useful supplementary data source to inform the estimation procedure. In addition to these two methodological criteria, two additional factors are also proposed. First, a methodology should be easily replicable to allow users to reproduce results with relative ease and understand at what stage (if any) erroneous estimates occur. Second, a degree of flexibility in the methodology is desirable. This can allow data users to apply the framework to different time periods or sets of countries.

Further desirable criteria, not listed in Table 1.1, may also be considered. Willekens (1994) suggested international migration flow data should be applicable to national demographic accounts. Hence, the net migration derived from the difference of the number of migrants received and sent in a single time period should be equal to the current population minus the population of the previous time period plus the natural change from births and deaths. Data users may also desire a methodology to estimate comparable migration flow data for migration by individual characteristics of migrants such as age and sex, to allow the further analyses of population behaviour and change. These additional properties are deemed beyond the scope of this thesis. The incorporation of international migration estimates into national demographic accounts would only be appropriate if an estimated migration table contained flows to and from all possible destinations in the world. However, this thesis is restricted to estimating aggregated flows between a selected set of nations.

The methodology developed in this thesis is applied to data from 15 countries in the European Union (EU) before the expansion of May 2004 (EU15). A series of tables, from 2002 to 2006 are studied. Larger sets of countries from alternative geographies such as the EU27 or EU31 are not studied in this thesis in order to provide a more concise illustration of reported data, estimates and the methodology. A concentration on European data is taken for a number of reasons. First, the study of international migration data in Europe is of growing importance due to the political reforms agreed by the European Parliament in 2004. These reforms have allowed citizens in the EU the right to move between, and reside freely in, member states (Kraler et al., 2006). Second, data from multiple countries in Europe are accessible from a number of international organizations, including Eurostat (the statistical office of the EU). Availability has been aided by policy makers of the European Parliament who 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. In 2007, Regulation No 862/07, obliged members to provide migration statistics which comply with a harmonized definition. Third, countries within EU vary in their population sizes and economic statuses but have relatively similar political structures. Measures of differences in the first two of these areas are more readily available, which will be of use for model based imputation methods. Fourth, recent European research projects such as Towards the Harmonisation of European Statistics on International Migration (THESIM) and MIgration MOdelling for Statistical Analyses (MIMOSA) have allowed differences between data collection methods and measurements used by national statistics institutes to be better understood. Finally, despite political reforms, regulations, similarities in member states and research reports, reported international migration flow data are still incomparable (see for example Nowok et al. (2006) or Kupiszewska and Nowok (2008)). National statistics institutes have struggled (and may continue) to adjust data collection and measurement procedures to provide data which is consistent across the region. In addition, there remain a number of countries which do not provide reported counts due to the lack of collection infrastructure or problems in the dissemination of data.

The reported data used in this thesis is obtained directly from the Eurostat web site, using origin-destination migration flows as supplied by sending and receiving countries (with local definitions of a migrant flow). This is further discussed in Chapter 4. Comparable data will be estimated according to the United Nations (UN) definition for a long term migration flow, i.e., the number of people who move to establish the usual place of residence in the destination country for twelve months or more (UN, 1998). This definition is also contained in EU regulations for the provision of international migration statistics by national statistics institutes (Giambattista and Poulain, 2006).

1.4 Thesis Structure

The study is structured in seven chapters. The following two chapters present known methodologies in statistics and international migration flow table estimation. New applications and extensions of the previously outlined methods are then presented in the remaining chapters of this thesis. Included in the early sections of these chapters are known statistical techniques, not previously introduced, as they require specific attention in the context of the study.

Chapter 2 introduces important statistical modelling techniques on which succeeding chapters will be heavily reliant. This commences with an introduction to the mathematical notation followed by an overview of generalized linear models, a unified modelling theory which can handle different response types. Included in this section are Poisson and log-normal regression models which are commonly used in modelling migration flows. The Iterative Reweighted Least Squares algorithm, a popular fitting method in statistical software for generalized linear models, is then outlined. This chapter concludes with a description of the negative binomial regression model, a useful extension to a Poisson regression model in the presence of overdispersion, but not a generalized linear model itself.

Chapter 3 reviews previous modelling frameworks used for estimating international migration flow tables. First, a brief outline of international migration flow tables and their data issues are given. This allows a basic understanding of the problems which motivated previous researchers to develop estimation techniques for European data. The subsequent sections concentrate on detailing frameworks for estimating migration flow tables, developed by Poulain (1993) and Raymer (2007). A comparison of the methods, with reference to the criteria in Table 1.1, are made and possible areas for extended study in following chapters are identified

Chapter 4 introduces a new methodological framework for the harmonization of international migration flow data. This uses constrained optimization techniques to estimate correction factors to scale reliable reported data. The first section presents international flow data for migration between EU15 countries, including background information and expert opinion on the characteristics of reported counts from each national statistic institute. This information helps inform the constrained optimization procedure to select data sources for which estimated correction factors are fixed to one, as they require no alteration. Data from unreliable sources are ignored as the scaling of reported flows will have no improving effect, and replacement values are estimated using missing data techniques outlined in the later chapters. Constrained optimization routines in statistical software are applied to minimize the difference in scaled reliable data. Estimates of harmonized flows are then calculated to obtain a set of incomplete international migration flow tables in each time period.

Chapter 5 uses the harmonized data of its preceding chapter to estimate missing migration flow data. It commences by reviewing models for migration flow tables. Negative binomial regression models are used in order to account for overdispersion in the data. A range of covariates on economic, demographic and geographical factors are considered. An appropriate model is selected based on the Akaike Information Criterion (AIC) statistics from the observed data. The model is then fitted by implementing the EM algorithm of Dempster et al. (1977) which accounts for missing data in the parameter estimation and imputes values for missing cells.

In Chapter 6, measures of variation for the imputations are obtained. The chapter commences by reviewing the convergence properties of the EM algorithm. An extension to the EM algorithm, the SEM algorithm of Meng and Rubin (1991) is then outlined. The SEM algorithm uses iterations in the EM algorithms to calculate the variance-covariance matrix of the parameters estimates and hence a measure of variability of imputations for previously missing data may be deduced. The succeeding section reviews the AICcd statistics (AIC for complete data) of Cavanaugh and Shumway (1998). The AICcd utilizes the SEM algorithm to allow the comparisons of models based on complete (both observed and missing) data, unlike in AIC statistics in Chapter 5.

Finally, Chapter 7 contains a summary of the findings, as well as the most important conclusions from the study. Together with a synopsis of the main results with reference to the criteria in Table 1.1, several recommendations for future research in the field of international migration flow tables are considered. The study is concluded by reflecting on the estimation framework in the context of international migration modelling and international migration data. The thesis is accompanied by an Appendix containing the S-Plus/R program codes used for the constrained optimization approach of Chapter 3, distance measures of Chapter 4, the EM algorithm in Chapter 5 and the SEM algorithm in Chapter 6.

Chapter 2

Statistical Modelling

2.1 Introduction

This chapter outlines the statistical modelling techniques for migration flow tables to set the stage for future chapters. It commences by introducing the notation to be used, demonstrated in a classic linear regression model. In the following section generalized linear models, a unified theory encompassing models for continuous, dichotomous and count responses are outlined. Included are Poisson and log-normal regression models which have often been used in modelling migration, as will be discussed further in Chapter 5. This is succeeded by general formulations for the mean and variance terms, likelihood equations and asymptotic variance-covariance matrix of parameters. These allow the Iterative Reweighted Least Squares algorithm, which is a popular fitting method in statistical software for generalized linear models, to be fully described. This algorithm is useful for finding maximum likelihood parameter estimates. The negative binomial regression model is then detailed. This model does not belong to the generalized linear model family but is a useful extension to a Poisson regression model in the presence of overdispersion. Included is a description of the Newton-Raphson method which is frequently used to fit parameters for the negative binomial regression model. This fitting method and associated extensions are also used in optimization problems, such as those implemented in Chapter 4. In the final section, techniques for estimating parameters in the presence of missing data are introduced.

2.2 Regression Models

In many scientific studies, interest lies in the relationship between two or more observable quantities. Regression analysis allows an estimation of the change in one quantity, y as a function of another, x. The quantity of primary interest in regression analysis, y is called the response variable. In a statistical model, the value of y is considered random in the sense that the observed values could have turned out differently due to the sampling process or natural variation of the population. Additional variables which are not considered as random, x are commonly known as covariates. In regression models, the conditional distribution of y given x is of interest, and studied in the context of a set of units $i = 1, \ldots n$, on which observed values y_i and x_i are measured. The set of p explanatory variables used in a model define the linear predictor which is commonly expressed as:

$$\beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tag{2.1}$$

where $\beta_1 \ldots \beta_p$ are the parameter estimates of the effect of y on $x_{i1} \ldots x_{ip}$. In many applications the variable x_{i1} is fixed at 1, so that $\beta_1 x_{i1}$ is constant for all i. The linear predictor of (2.1) may also be written in matrix notation as $\mathbf{x}_i^T \boldsymbol{\beta}$ or a component of $\mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1 \ldots \beta_p)^T$, $\mathbf{x}_i^T = (x_{i1} \ldots x_{ip})$ and $\mathbf{X}$ is a $n \times p$ matrix. In a classic linear regression model the response variable is fully described by specifying the conditional probability density of y given the linear predictor,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.2}$$

where $\mathbf{y} = (y_1 \ldots y_n)^T$ are continuous responses and the residuals $\mathbf{u} = (u_1 \ldots u_n)^T$ are independently normally distributed with zero mean and constant variance σ^2 for all units, $u_i \sim N(0, \sigma^2)$. The classic linear regression model can be alternatively defined by setting the conditional expectation of the responses, μ_i, given the linear predictor, equal to $\mathbf{x}_i^T \boldsymbol{\beta}$:

$$\mu_i \equiv E(y_i | \boldsymbol{\beta}, \sigma, \mathbf{x}_i^T) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{2.3}$$

where y_i are independent normally distributed with mean μ_i and variance σ^2.

2.3 Generalized Linear Models

Linear regression models are part of a range of statistical models, known as generalized linear models (Nelder and Wedderburn, 1972). These models link together a variety of random response variables to a systematic linear predictor. This includes models where the assumption of a linear relationship or normal variations of a response variable may not be appropriate, such as a log-linear relationship or a Poisson count response, both of which will be expanded upon in this section. Agresti (2002, p116-7) outlines three components used in the specification of generalized linear models:

(a) A random component identifying the natural parameter, θ_i, where the distribution of the response variable is a member of the natural exponential family. A natural exponential distribution has probability density function of the form;

$$f(y_i | \theta_i, \phi) = \exp\left\{ \frac{\theta_i y_i - a(\theta_i)}{c(\phi)} + b(y_i, \phi) \right\}, \tag{2.4}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are functions depending on the distribution. The value of θ_i may vary for units $i = 1, \ldots, n$, depending on the values of the explanatory variables. The dispersion parameter ϕ is equal to unity for some distributions.

(b) A systematic component to relate the vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$ to the explanatory values, through a linear model, using the linear predictor, $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

(c) The link function, $g(\cdot)$, to connect the random and systematic components. When μ_i is the conditional expectation of the response, a generalized linear model links μ_i to η_i by $\eta_i = g(\mu_i)$. As $g(\cdot)$ is a monotonic differentiable function, it can be expressed in terms of explanatory variables by $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$.

These three components transform the mean of a response variable, in the natural exponential family distribution, from a non-linear to a linear model.

2.3.1 Normal and Log-Normal Distribution

In a continuous case, the response variable can be assumed to be independently normally distributed with parameters (μ_i, σ^2) for the mean and variance respectively. The probability density function of this distribution is given by

$$f(y_i|\mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2}. \tag{2.5}$$

In the generalized linear model format we can re-express the probability density function in the representation of the natural exponential family of (2.4):

$$f(y_i|\theta_i, \sigma^2) = \exp\left\{ \frac{\mu_i y_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2} \right\}, \tag{2.6}$$

where $a(\theta_i) = \frac{\theta_i^2}{2}$, $b(y_i, \phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}$, $c(\phi) = \sigma^2$ and $\theta_i = \mu_i$. Hence, the expectation of the random variable in the generalized linear model format is μ_i and we connect this to the systematic component using the (canonical) identity link function $g(\mu_i) = \mu_i$. This gives the regression model of (2.3);

$$\eta_i = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{2.7}$$

A linear regression model may be applied in a generalized linear model framework when the response is non-linear. When the assumption of normality does not hold a log link function rather than the identity link of (2.7) is commonly used. The log-normal distribution is typically assumed for response variables which take positive values on the continuous scale, where there exist no theoretical possibility of a non-positive value occurring. A traditional approach to modelling data that has log-normal distribution is to normalize the response in (2.7), relative to the linear predictor, by calculating the logarithm of each unit's outcome. Hardin and Hilbe (2001, p59) noted that this method leads to an inconvenient interpretation of fitted values and parameter estimates which are in terms of a log response. They suggest an alternative approach is to internalize within the model itself the log transformation of the response. This can be represented in the form of the natural exponential family,

$$f(y_i|\theta_i, \sigma^2) = \exp\left\{ \frac{\log(\mu_i) y_i - \frac{(\log(\mu_i))^2}{2}}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2} \right\}, \tag{2.8}$$

where $a(\theta_i) = \frac{(\log(\theta_i))^2}{2}$, $b(y_i, \phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}$, $c(\phi) = \sigma^2$ and $\theta_i = \log(\mu_i)$. Hence, the log link function $g(\mu_i) = \log(\mu_i)$ is used to connect the random and systematic components giving a log-linked normal regression model

$$\eta_i = \log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{2.9}$$

2.3.2 Poisson Distribution

In a discrete case, a response variable of count data can be assumed to have a Poisson distribution with rate parameter μ. The probability density function of this distribution is given by

$$f\left(y_i \left| \mu_i \right.\right) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots. \tag{2.10}$$

In a generalized linear modelling format we can re-express the probability density function in the representation of the natural exponential family of (2.4):

$$f\left(y_i \left| \theta_i \right.\right) = \exp\left\{ \frac{y_i \log \mu_i - \mu_i}{1} - \log y_i! \right\}, \tag{2.11}$$

where $a\left(\theta_i\right) = e^{\theta_i}$, $b\left(y_i, \phi\right) = -\log y_i!$, $c\left(\phi\right) = 1$ and $\theta_i = \log \mu_i$. The expectation of the random variable in the generalized linear model format is μ_i, and is connected to the systematic component using the log link function $g\left(\mu_i\right) = \log \mu_i$. This results in a Poisson regression model,

$$\eta_i = \log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{2.12}$$

which is identical to the model expressed in (2.9), but where the response is no longer assumed to be log-normal.

When modelling count data, it is often of interest to measure the rate at which the count occurs rather than the count itself. The rate can be obtained by dividing the count by the related exposure, e_i, of each unit. Hence, a Poisson rates model can be derived from (2.12) as

$$\begin{aligned} \log\left(\frac{\mu_i}{e_i}\right) &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \log(\mu_i) &= \log(e_i) + \mathbf{x}_i^T \boldsymbol{\beta} \qquad (2.13) \end{aligned}$$

where $\log(e_i)$ is a known offset term. Poisson regression models with offset terms have been used in a previous framework for the estimation of international migration such as Raymer (2007). This will be outlined in Chapter 3 of this thesis.

2.4 Fitting Generalized Linear Models

Maximum likelihood estimates are frequently used in migration models as they posses very desirable asymptotic properties such as consistency, asymptotic normality and asymptotic robustness (Sen and Smith, 1995, p457-69). In generalized linear models, such estimators are found within most statistical software packages using the Iteratively Reweighted Least Squares (IRLS) procedure, which McCullagh and Nelder (1989, p41-2) proved to converge to the maximum likelihood solutions. In this section the IRLS fitting method will be described after some necessary properties of generalized linear models are outlined.

2.4.1 Mean and Variance

The mean and variance of the random component in a generalized linear model may be obtained in a general form, allowing the maximum likelihood estimates to be found using IRLS. Assuming the responses of all units are independent, the likelihood for a generalized linear model is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi). \tag{2.14}$$

where $\boldsymbol{\theta}$ is the p-dimensional parameter vector. If we let l_i denote the log-likelihood for the i^{th} observation then

$$l(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log f(y_i|\theta_i,\phi). \tag{2.15}$$

Therefore, using the probability distribution of the natural exponential family expressed in (2.4) we may deduce by differentiation

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - a'(\theta_i)}{c(\phi)} \\ \frac{\partial^2 l_i}{\partial \theta_i^2} &= -\frac{a''(\theta_i)}{c(\phi)}, \end{aligned} \tag{2.16}$$

where $a'(\theta_i)$ and $a''(\theta_i)$ are the first and second derivatives of $a(\cdot)$ evaluated at θ_i. As Cox and Hinkley (1974) showed, the general likelihood results: $E\left(\frac{\partial l_i}{\partial \theta_i}\right) = 0$ and $-E\left(\frac{\partial^2 l_i}{\partial \theta_i^2}\right) = E\left(\frac{\partial l_i}{\partial \theta_i}\right)^2$ are satisfied by the natural exponential family, so

$$E\left(\frac{y_i - a'(\theta_i)}{c(\phi)}\right) = 0$$

and hence

$$E(y_i) = a'(\theta_i), \tag{2.17}$$

for the mean, and

$$E\left(\frac{a''(\theta_i)}{c(\phi)}\right) = E\left(\frac{y_i - \mu_i}{c(\phi)}\right)^2$$

thus

$$(E(y_i) - \mu_i)^2 = \frac{a''(\theta_i)(c(\phi))^2}{c(\phi)}$$

and

$$\mathrm{Var}(y_i) = a''(\theta_i)c(\phi), \tag{2.18}$$

for the variance. Hence, the mean and variance of any distribution from the natural exponential family can be easily derived from (2.17) and (2.18). For example, for the Poisson distribution, where $a(\theta_i) = e^{\theta_i} = \mu_i$ as $\theta_i = \log \mu_i$, the mean $E(y_i) = a'(\theta_i) = \mu_i$ and variance $\mathrm{Var}(y_i) = a''(\theta_i) \times 1 = \mu_i$.

2.4.2 Likelihood Equations

In order to obtain maximum likelihood parameter estimates for a generalized linear model we must first obtain the likelihood equations. Assuming that the responses of n units are independent, the likelihood for generalized linear model can be expanded from (2.15) as

$$l(\boldsymbol{\beta}|y) = \sum_{i=1}^{n} \frac{\theta_i y_i - a(\theta_i)}{c(\phi)} + \sum_{i=1}^{n} b(y_i, \phi), \tag{2.19}$$

where $l(\boldsymbol{\beta}|y)$ reflects the dependence of $\boldsymbol{\theta}$ on the model parameters. The likelihood equations can then be derived by differentiating the log-likelihood with respect to an arbitrary β_j, using the chain rule, and then equating to zero:

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = 0. \tag{2.20}$$

As $\mu_i = E(y_i) = a'(\theta_i)$,

$$\begin{aligned}
\frac{\partial l}{\partial \theta_i} &= \frac{y_i - a'(\theta_i)}{c(\phi)} = \frac{y_i - \mu_i}{c(\phi)} \\
\frac{\partial \theta_i}{\partial \mu_i} &= \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \frac{1}{a''(\theta_i)} = \frac{c(\phi)}{\mathrm{Var}(y_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} &= g'(\mu_i), \text{ as } \eta_i = g(\mu_i) \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}, \text{ as } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta},
\end{aligned} \tag{2.21}$$

where x_{ij} is the (i,j) element of $\mathbf{X}$ and $g'(\mu_i)$ is the first derivative of the link function $g(\cdot)$ evaluated at μ_i. We may substitute these expressions into the likelihood equations to give

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{c(\phi)} \frac{c(\phi)}{\mathrm{Var}(y_i)} g'(\mu_i) x_{ij} = 0 \\
&= \sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{\mathrm{Var}(y_i)} g'(\mu_i).
\end{aligned} \tag{2.22}$$

The likelihood equations for any distribution from the natural exponential family can be directly obtained from (2.22). For example, a normal distributed response in a classic linear regression model has $\mathrm{Var}(y_i) = \sigma^2$ and $g'(\mu_i) = 1$. Hence, the likelihood equations are

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{\sigma^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) = 0. \tag{2.23}$$

2.4.3 Asymptotic Variance-Covariance Matrix of Parameters Estimates

The asymptotic variance-covariance matrix for parameter estimates is required to provide a useful simplification in the IRLS procedure. This may be derived from the inverse of the $p \times p$ Fisher (or expected) information matrix $\mathbf{I}(\boldsymbol{\beta})$, which has elements $(\mathbf{I}(\boldsymbol{\beta}))_{jk} = -E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right)$ for any two arbitrary parameters. Using the general likelihood results of

Cox and Hinkley (1974) we may express $\mathbf{I}(\boldsymbol{\beta})$ as

$$\begin{aligned} -E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) &= E\left(\frac{\partial l_i}{\partial\beta_j}\frac{\partial l_i}{\partial\beta_k}\right) \\ &= E\left\{\frac{x_{ij}(y_i-\mu_i)}{\mathrm{Var}(y_i)}g'(\mu_i)\frac{x_{ik}(y_i-\mu_i)}{\mathrm{Var}(y_i)}g'(\mu_i)\right\} \\ &= E\left\{\frac{x_{ij}x_{ik}(y_i-\mu_i)^2}{\mathrm{Var}(y_i)^2}g'(\mu_i)^2\right\} \\ &= \frac{x_{ij}x_{ik}}{\mathrm{Var}(y_i)}g'(\mu_i)^2 \\ &= \frac{x_{ij}x_{ik}}{c(\phi)V(\mu_i)}g'(\mu_i)^2, \qquad (2.24) \end{aligned}$$

where $E(y_i-\mu_i)^2 = \mathrm{Var}(y_i) = V(\mu_i)c(\phi)$ and $V(\mu_i)$ is the variance function evaluated at μ_i. Since $l(\boldsymbol{\beta}) = \sum l_i$

$$-E\left(\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{c(\phi)V(\mu_i)}g'(\mu)^2, \qquad (2.25)$$

which can be generalized to matrix formation for $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a $p\times p$ diagonal matrix with main diagonal elements

$$w_i = \frac{g'(\mu_i)^2}{c(\phi)V(\mu_i)}. \qquad (2.26)$$

The asymptotic variance-covariance matrix, $\mathbf{V}$, of the parameters estimates, $\widehat{\boldsymbol{\beta}}$, is estimated by

$$\mathbf{V} \;= \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1} {=} (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}, \qquad (2.27)$$

where $\widehat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\widehat{\boldsymbol{\beta}}$. The asymptotic variance-covariance matrix of any distribution from the natural exponential family can be directly derived from (2.26). For example, a Poisson regression model has $c(\phi) = 1, V(\mu_i) = \mu_i$ and $g'(\mu_i) = \mu_i$. Hence $\mathbf{W}$ in (2.27) has main diagonal elements $w_i = \frac{\mu_i^2}{\mu_i} = \mu_i$.

2.4.4 Iterative Reweighted Least Squares

For the likelihood equations of a classic linear regression model the maximum likelihood estimators of $\boldsymbol{\beta}$ can be found by re-expressing (2.23) for $\boldsymbol{\beta}$, in a matrix notation:

$$\mathbf{X}\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0$$

hence,

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \qquad (2.28)$$

where $\mathbf{W}$ is a diagonal matrix with diagonal elements equal to σ^2. As the residual variance is homoscedastic with all diagonal elements of $\mathbf{W}$ equal, this term can be dropped leaving an ordinary least squares estimate for $\boldsymbol{\beta}$. If the residuals are heteroscedastic, where $c(\phi) = \sigma_i^2$, then the maximum likelihood estimate of $\boldsymbol{\beta}$ can be found using a weighted least squares estimator, as given in (2.28) with weights σ_i^2, as $c(\phi) = \sigma_i^2, V(\mu_i) = 1$ and $g'(\mu_i) = 1$ in (2.26). Hence the likelihood equations are linear in $\boldsymbol{\beta}$

$$\frac{\partial l}{\partial\beta_j} = \sum_{i=1}^{n}\frac{x_{ij}}{\sigma_i^2}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) = 0. \qquad (2.29)$$

When the likelihood equations are non-linear, as in some generalized linear models, we may use the IRLS algorithm to estimate $\boldsymbol{\beta}$. The algorithm works by linearizing the likelihood equations for the application of weighted least squares at each cycle of the iterative procedure. Each iteration cycle, r, estimates a current iterate, $\boldsymbol{\beta}^r = (\beta_i^r, \ldots, \beta_p^r)^T$ with corresponding mean, μ_i^r, and a working variate, z_i^r, where

$$z_i^r = \mathbf{x}_i^T \boldsymbol{\beta}^r + (y_i - \mu_i^r) g'(\mu_i^r), \tag{2.30}$$

hence

$$y_i - \mu_i^r = \frac{(z_i^r - \mathbf{x}_i^T \boldsymbol{\beta}^r)}{g'(\mu_i^r)}. \tag{2.31}$$

Estimates of $\boldsymbol{\beta}^r$ can be updated using weighted least squares, from the working variate vector $\mathbf{z} = (z_1, \ldots, z_n)^T$,

$$\boldsymbol{\beta}^{r+1} = (\mathbf{X}'\mathbf{W}^r\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^r\mathbf{z}^r, \tag{2.32}$$

where the diagonal elements of $\mathbf{W}^r$ are given by $g'(\mu_i^r)^2 c(\phi) V(\mu_i)$. When this weight is incorporated into (2.29) as if the model were linear, the likelihood equations,

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i^r)^2 c(\phi) V(\mu_i^r)} (z_i^r - \mathbf{x}_i^T \boldsymbol{\beta}^r) \\ &= \sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i^r) c(\phi) V(\mu_i^r)} \frac{(z_i^r - \mathbf{x}_i^T \boldsymbol{\beta}^r)}{g'(\mu_i^r)} \\ &= \sum_{i=1}^{n} \frac{x_{ij}}{g'(\mu_i^r) c(\phi) V(\mu_i^r)} (y_i - \mu_i^r), \end{aligned} \tag{2.33}$$

are the same as the original weighted least squares likelihood equations, (2.22), except the weights are fixed at the estimates from the previous iteration, r. Hence, solving these equation using weighted least squares of (2.32) gives estimates of $\boldsymbol{\beta}^{r+1}$ which may lead to new calculated weights, then new estimates using reweighted least squares and so on, iterating until convergence. Skrondal and Rabe-Hesketh (2004, p192) noted that this method is identical to Fisher Scoring, an alternative iterative method for solving likelihood equations.

Generalized linear models are commonly applied by social scientists to model migration data, as will be discussed in Chapter 5. This is often undertaken to gain a substantive understanding of movements. In this thesis, the aspects discussed in this section have an alternative use, to derive estimates for migration data that is currently unreliable. In Chapter 5 the IRLS procedure is used within the Expectation Maximization (EM) algorithm for fitting migration models to incomplete data. This allows imputations for missing values to be obtained. In Chapter 6 the IRLS procedure is used within the Supplemented EM algorithm to derive estimates of the asymptotic variance-covariance matrix for model parameter estimates in the presence of missing data.

2.5 Negative Binomial Regression Models

The negative binomial distribution has two-parameters that allow a mean and variance to be fitted separately, as opposed to a single parameter Poisson regression model. It may be considered in two ways: as a marginal distribution of a Poisson random variable where the rate parameter has a gamma distribution or as a probability function in its own right for the observation of y failures before a nth successes in a series of Bernoulli trials. Hardin and Hilbe (2001, p140) noted that when considered as the first of these approaches (as will be the case in this thesis), the negative binomial distribution is not a member of the exponential family and hence cannot be considered in the generalized linear model framework. Considered as such the probability density function can be expressed as

$$f(y_i|\mu_i,\alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}, \quad y_i = 0,1,\ldots, \quad \alpha \geq 0, \tag{2.34}$$

such that

$$\mu_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \tag{2.35}$$

where μ_i is the mean of a Poisson distribution and α is a dispersion parameter. The regression model of a negative binomial response takes the same format as the Poisson regression model of (2.12). The model has a mean of μ_i and variance function of $\mu_i + \alpha\mu_i^2$. When the overdispersion is zero the Poisson model is obtained.

2.5.1 Asymptotic Variance Covariance Matrix

Cameron and Trivedi (1998, p71) showed that for the negative binomial regression model the maximum likelihood estimates are the solution to the first order conditions

$$\sum_{i=1}^{n}\frac{y_i-\mu_i}{1+\alpha\mu_i}\mathbf{x}_i^T = 0. \tag{2.36}$$

Hence, the asymptotic variance-covariance matrix for the parameter estimates can be derived as

$$\begin{bmatrix} E\left(\frac{\partial^2 l_i}{\partial\beta_j\beta_k}\right)^{-1} & 0 \\ 0 & E\left(\frac{\partial^2 l_i}{\partial\alpha^2}\right)^{-1}\end{bmatrix}, \tag{2.37}$$

where the elements off the block diagonal are solutions to $\frac{\partial^2 l_i}{\partial\beta_j\partial\alpha} = 0$ for each j. Thus the variance-covariance between elements of the parameter vector $\boldsymbol{\beta}$ are the same as (2.27) for a generalized linear model, where $c(\phi) = 1, V(\mu_i) = \mu_i+\alpha\mu_i^2$, $g'(\mu_i) = \mu_i$, and $\mathbf{W}$ has main diagonal elements $w_i = \frac{\mu_i^2}{\mu_i+\alpha\mu_i^2} = \frac{\mu_i}{1+\alpha\mu_i}$. Cameron and Trivedi (1998, p72) demonstrated that by expressing $\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})} = \prod_{g=0}^{y_i-1}(g+\alpha^{-1})$ in (2.34), the variance of α in (2.37) can be deduced as

$$\left[\sum_{i=1}^{n}\frac{1}{\alpha^4}\left(\log(1+\alpha\mu_i) - \sum_{g=1}^{y_i-1}\frac{1}{(g+\alpha^{-1})}\right)^2 + \frac{\mu_i}{\alpha^2(1+\alpha\mu_i)}\right]^{-1}. \tag{2.38}$$

2.5.2 Fitting Negative Binomial Regression Model

Agresti (2002, p560-1) noted that a negative binomial model may be fitted in a similar manner as Poisson regression models when the dispersion parameter is known. This can be implemented using the IRLS procedure. When the dispersion parameter is not known, three possible methods exist to obtain maximum likelihood parameter estimates: a Newton-Raphson routine for fitting all parameters simultaneously; the evaluation of the profile likelihood for various fixed α, and an alternation strategy of 1) using IRLS to solve mean parameter estimates $\boldsymbol{\beta}$, for fixed α and 2) using Newton-Raphson to estimate α from fixed $\boldsymbol{\beta}$, until convergence.

The Newton-Raphson method (also known as an Newton optimizer) is an iterative routine for finding roots in non-linear equations, in one or more dimensions. It can be applied to likelihood functions to find local maxima and local minima often with rapid convergence. The method is also used for other optimization problems in non-statistical settings. The Newton-Raphson method considers an approximation of the derivatives of the likelihood function, using a first order Taylor expansion around a parameter estimate θ:

$$L(\theta + \Delta\theta) = L(\theta) + L'(\theta)\Delta\theta + \frac{1}{2}L''(\theta)(\Delta\theta)^2, \tag{2.39}$$

This expression attains its extremum when $\Delta\theta$ solves the linear equation

$$L'(\theta) + L''(\theta)\Delta\theta = 0, \tag{2.40}$$

and $L''(\theta)$ is positive. Thus, provided that $L(\theta|y)$ is a twice-differentiable function and the initial guess of a working estimate, θ^r, is chosen close enough to the stationary point, θ^*, then

$$\theta^{r+1} = \theta^r - \frac{L'(\theta^r)}{L''(\theta^r)}, \tag{2.41}$$

will converge towards θ^*. When fitting a negative binomial regression model, this method can be used to estimate the dispersion parameter, where $\theta = \alpha$ in (2.41) and current estimates of mean parameters, $\boldsymbol{\beta}$ are provided by IRLS. The asymptotic variance-covariance matrix of (2.37) may then be fully obtained given the estimate of α for (2.38).

The Newton-Raphson routine can be generalized to several dimensions for multiple parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Replacing the derivative of (2.41) with the p dimensional gradient vector, $\mathbf{v}^T = \left(\frac{\partial L(\theta_1)}{\partial \theta_1}, \ldots, \frac{\partial L(\theta_p)}{\partial \theta_p}\right)$, and the reciprocal of the second derivative with the inverse of the Hessian matrix, $\mathbf{H}$, where an element $h_{jk} = \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}$. Hence, a modified iterative scheme for multiple parameters is obtained:

$$\boldsymbol{\theta}^{r+1} = \boldsymbol{\theta}^r - (\mathbf{H}^r)^{-1}\mathbf{v}^r. \tag{2.42}$$

For a negative binomial regression model this generalized routine can also be used to estimate all parameter, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)$.

In comparison with generalized linear models, negative binomial regression models have been more limited in their application to migration data by social scientists. This will be discussed further in Chapter 5 where negative binomial regression models are used to obtain imputations for missing data. In Chapter 4, the quasi-Newton optimizer (related to (2.42) with numerical estimates of $\mathbf{v}$ and $\mathbf{H}$) will be used in a non-statistical setting to estimate correction factors that minimize the distance between reported migration flow counts, subject to a set of constraints.

2.6 Statistical Modelling of Missing Data

International migration flow data is often missing. In order to estimate missing flows, model based methods may be used to derive parameter estimates that account for data incompleteness. In this thesis, the EM algorithm of Dempster et al. (1977) is used to find maximum likelihood estimates of model parameters in the presence of missing data. In Chapter 6, the Supplemented EM (SEM) algorithm of Meng and Rubin (1991) is used to estimate the variance covariance matrix of parameter estimates when there is missing data. Within these algorithms, methods described in this chapter, such as IRLS and the Newton-Raphson routine are used repetitively. Unlike the fitting methods discussed in this chapter, standard functions for these algorithms are unavailable in statistical software. To this end, S-Plus/R functions were written to fit negative binomial regression models for international migration flow data using the EM and SEM algorithms. A full discussion of the algorithms, their properties and the functions written are hence deferred until the appropriate chapters of this thesis.

Chapter 3

A Review of Methodologies for Estimating an International Migration Flow Table of Comparable Data

3.1 Introduction

At present, the responsibility for the collection of international migration flow data rests with individual national statistics institutes. Consequently, data on a considered flow can be missing, reported by the sending country, reported by the receiving country or reported by both the sending and receiving country. For the last situation, in which two sources of information are possible for one particular flow, data rarely match due to differences in collection and measurement.

In this chapter, previous frameworks to estimate international migration flow tables of comparable data are reviewed. The strengths, weaknesses and differences of methodologies are made with reference to the desirable criteria shown in Table 1.1. The next section commences with an outline of the problems in the comparability of international migration flow data. The succeeding section provides an introduction to migration flow tables, illustrated with generated counts. This allows a clear presentation of data issues and frameworks for estimating flow values. The following sections concentrate on different frameworks for estimating migration flow tables in a single year. The first, developed by Poulain (1993), used a constrained optimization to estimate correction factors to scale reported data. This method, and more recent extensions, are illustrated with an example. The second methodology, initially introduced by Raymer (2007) modelled the components of a saturated regression model applied to a migration flow table. Both methodologies are discussed separately, before concluding with a comparative review.

3.2 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 statistics 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 and Nowok (2008). The incomparability between data sources in any time period is 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.

3.2.1 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 statistics institutes.

Data collection methods may influence the completeness and accuracy of reported migration flows (Nowok et al., 2006). National statistics 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 data bases 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 Great Britain the International Passenger Survey (IPS) is used to provide international migration flow data. In order to supply sufficient detail for analysis, the sample size must be very large, otherwise unexpected irregularities appear for specific origin-destination flows (Perrin and 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 and 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 often unknown, only that they lived outside the country one-year or five years prior to the census or survey date. For some data sources the intended duration, rather than the actual duration is used. Under an actual duration measure, reporting of counts are delayed to allow the period used in the timing criteria to pass, whereas under an intended duration measure an assumption that the intended period will become the actual duration is made. Nowok et al. (2006) noted that some national statistics institutes measure intended duration measures for non-national immigrants by the period specified in the authorization to stay, which may hence differ from the actual duration.

The coverage of difficult to measure population groups, such as asylum seekers, students and illegal residents 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. Among the EU member states, only Spain allows the registration of illegal migrants through a pardon system (Breem and Thierry, 2006b), allowing the capture of data on this difficult to measure group.

3.2.2 Data Dissemination Methods

Differences in the dissemination of migration flow statistics can be derived from alternative methods for handling migrants with unknown origins or destinations and limitations on the collection of specific flow information.

National statistics institutes may struggle to fully disseminate information on the origin or destination of migrants. 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 is relatively large with regard to the total migration count. For other nations, this count may be small or zero. Hence, when comparing differences in reported migration flows between multiple nations, the counts of movements associated with unknown origins or destinations must be considered.

Migration flow data may be completely available, partially available or completely unavailable. Partial availability can occur for data from countries that have a domestic need to measure only certain flows. For example, in 2002 Ireland produced estimates of total movements to and from only three areas: Great Britain, 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 to and from Great Britain are unable to provide estimates for individuals origins or destinations where low volumes of movements exist (Perrin and Poulain, 2006b). For some countries, no migration flow data may be produced. For data sources from member states 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 and Thierry, 2006a). Conversely, similar sized countries, such as Italy and Germany, 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 might be caused by national statistics institutes not providing, or the organizations not publishing data, despite collection procedures being in place.

3.3 International Migration Flow Tables

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 entries in the migration flow table (which corresponds to 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 statistics 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. A simple example of such tables is shown in Table 3.1, where data is generated using a Poisson random process (from the `rpois` function in S-Plus 6.2) with rate parameter equal to 10. Data were not generated for migrants received and sent by region E.

Table 3.1: Simulated Migration Flow Tables from Receiving (left) and Sending (Right) Countries

Origin	Destination					Total
	A	B	C	D	E	
A		8	12	11		31
B	5		7	8		20
C	11	8		5		24
D	12	10	17			39
E	12	7	10	7		36
Total	40	33	46	31		150

Origin	Destination					Total
	A	B	C	D	E	
A		7	10	8	11	36
B	7		10	8	6	31
C	5	4		14	9	32
D	11	11	14		15	51
E						
Total	23	22	34	30	41	150

In both tables the origins are shown on the vertical axis and destinations on the horizontal axis. Data collected by the receiving destination countries in the left hand table form a vertical pattern. In the same manner, the origin reported values in the right hand table, as collected by the sending nations, form a horizontal pattern.

For any single year, demand exists for a single table with one comparable flow value for each origin-destination combination. As discussed in the previous section, data for international migration flow tables often lack comparability (in the same manner as Table 3.1). Sources of the incomparability between a flow reported in each table can be attributed to two areas: inconsistencies and incompleteness.

Where two sources of information exist for one particular flow, the data might or might not resemble each other because of differences in definitions and collection systems. These differences result in data inconsistencies similar to those shown in Table 3.1. For example, the flows from country A to B are very similar (8 and 7 respectively), whereas the flows from country C to B are very different (8 and 4 respectively). Inconsistencies in reported flow values create a confusing impression as to which data source (if any) is to be preferred. In turn, more doubts are apparent when considering values in cells where only a single value is reported (such as from country A to E). Values compared across columns for receiving data or across rows for sending data could be higher or lower depending on definitions and data collection methods rather than more or less migrants entering or leaving countries. Where both reported flow counts are missing data users are unable to obtain any idea of the level of flows between nations. Together, problems of inconsistent and incomplete data make comparisons of migration flows across a set of countries difficult.

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 has a significant influence. Previous methodologies for adjusting and imputing missing data have been created to address this demand. These have tended to be broken into multiple stages, addressing the problems of inconsistencies and incompleteness through a mixture of methods.

3.4 Constrained Optimization

Estimates of a complete migration flow table between 28 European nations in 2004 were calculated by Poulain and Dal (2007) (and Poulain and Dal (2008)) as part of the MIMOSA project. This involved decomposing the estimation of international migration flow tables into three stages: harmonization of referee countries data using a constrained optimization, estimation of flows between referee and non-referee countries and estimation of flows between non-referee countries. Refereed countries were chosen according to the availability of flow data and expert judgement on their reliability. Full details on this decision are left to the discussion section. In this section, the methodology of Poulain and Dal (2008) is initially discussed mathematically, followed by an illustrated example from the hypothetical data presented in Table 3.1.

The first step of the procedure of Poulain and Dal (2008) built on earlier work (Poulain, 1993, 1999) which used smaller migration tables of flows between selected countries without missing data. The estimation of harmonized values in these studies required an underlying assumption that differences in the reported counts of flows between countries are fixed. Thus the distance between counts represent the non-random discordance in the collection and measurement of migration flows between any two national statistics institutes. Under this assumption the equality

$$r_j m_{ij1} = s_i m_{ij2}, \tag{3.1}$$

is believed to hold, where r_j, scales receiving data reported in destination j, and s_i scales sending data reported in origin i. When correction factors are unknown, Poulain (1993) suggested that they can be estimated by minimizing the Euclidean distance,

$$f(r_j, s_i|m_{ijk}) = \sum_{i,j}(r_j m_{ij1} - s_i m_{ij2})^2. \tag{3.2}$$

Poulain and Dal (2007) proposed the replacement of this measure with a Chi-Squared distance function to allow the sum of differences in adjusted cell counts to be weighted by the observed data,

$$f(r_j, s_i|m_{ijk}) = \sum_{i,j} \frac{(r_j m_{ij1} - s_i m_{ij2})^2}{m_{ij1} + m_{ij2}}. \tag{3.3}$$

Estimates for the correction factors from both distance measures can be obtained by finding the root of the partial differential equation. The optimal solution to these equations are for all correction factors to equal zero. In order to determine non-zero parameter values, Poulain and Dal (2007) imposed a constraint,

$$c(\boldsymbol{\theta}|m_{ijk}) = \sum_{i,j} \frac{r_j m_{ij1} + s_i m_{ij2}}{2} = \sum_{i,j} \max(m_{ij1}, m_{ij2}). \tag{3.4}$$

Earlier studies such as Poulain (1993) used alternative constraints, based on the reported receiving data. In order to minimize (3.3) with respect to this constraint the method of Lagrange multipliers was used. This can be illustrated by letting the distance and constraint functions be denoted by $f(\boldsymbol{\theta}|m_{ijk})$ and $c(\boldsymbol{\theta}|m_{ijk})$ for the parameter set $\boldsymbol{\theta} = (\mathbf{r}, \mathbf{s})$, where $\mathbf{r}$ and $\mathbf{s}$ are the sets of receiving and sending correction factors for referee data sources respectively. The method of Lagrange multipliers achieves the stationary points of $\boldsymbol{\theta}$ by setting the partial differentials of

$$L(\boldsymbol{\theta}, \lambda|m_{ijk}) = f(\boldsymbol{\theta}|m_{ijk}) - \lambda c(\boldsymbol{\theta}|m_{ijk}), \tag{3.5}$$

to zero where λ is the Lagrange multiplier. For the Chi-Squared distance function and constraint used in Poulain and Dal (2007), the partial derivatives for the parameter set and Lagrange multiplier:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta}|m_{ijk})}{\partial r_j} &= \sum_{i,j} \frac{2m_{ij1}(r_j m_{ij1} - s_i m_{ij2})}{m_{ij1} + m_{ij2}} - \lambda \sum_{i,j} \frac{m_{ij1}}{2} = 0 \\
\frac{\partial L(\boldsymbol{\theta}|m_{ijk})}{\partial s_i} &= \sum_{i,j} \frac{-2m_{ij2}(r_j m_{ij1} - s_i m_{ij2})}{m_{ij1} + m_{ij2}} - \lambda \sum_{i,j} \frac{m_{ij2}}{2} = 0 \\
\frac{\partial L(\boldsymbol{\theta}|m_{ijk})}{\partial \lambda} &= -\sum_{i,j} \frac{r_j m_{ij1} +_i m_{ij2}}{2} + \sum_{i,j} \max(m_{ij1}, m_{ij2}) = 0,
\end{aligned} \tag{3.6}$$

are all linear equations. These can be represented by a system of equations in matrix format as such,

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}, \tag{3.7}$$

where $\mathbf{A}$ includes the constant terms expressed in (3.6), $\boldsymbol{\theta} = (\mathbf{r}, \mathbf{s}, \lambda)^T$ and $\mathbf{b}$ is a vector of zeros, with except the last element, which is equal to $\sum_{i,j} \max(m_{ij1}, m_{ij2})$.

Poulain and Dal (2008) suggested that in order to harmonize values to a known definition the parameter set should be normalized to a selected country. This allows correction factors in $\boldsymbol{\theta}$ to be interpreted as the effect of different measurement and collection systems in each data source in reference to the selected (normalized) data source. In their demonstrated example, this was implemented with Swedish receiving data, which were highly regarded by data experts due to the collection methods and definitional measure used (Herm, 2006a). Dividing all parameter estimates by the estimated parameter corresponding to Swedish receiving data, resulted in the constraint of (3.4) no longer holding.

In the second and third stage, correction factors for all remaining data sources are estimated by dividing the scaled flow values estimated using the correction factors from the first stage with the original reported data,

$$\begin{aligned} r_{j'} &= \frac{\sum_i s_i m_{ij'2}}{\sum_i m_{ij'1}} \\ s_{i'} &= \frac{\sum_j r_j m_{i'j1}}{\sum_j m_{i'j2}}, \end{aligned} \tag{3.8}$$

where i' and j' represent the respective row and columns corresponding to non-refereed countries. When no original reported data exists for the estimation of the correction factors in (3.8), alternative migration data such as origin-destination migrant stocks or migration flows defined by country of citizenship (rather than country of previous/next residence) are imputed.

Final estimated flow values, y_{ij}, for the migration flows from origin i to destination j, are derived by using the set of correction factors for both refereed and non-refereed countries to scale the reported (or imputed) data according to a set of preferences,

$$y_{ijt} = \begin{cases} \frac{1}{2}(r_j m_{ij1} + s_i m_{ij2}) & \text{if } r_j \text{ and } s_i \text{ exist} \\ \frac{1}{2}(r_j m_{i'j1} + s_{i'} m_{i'j2}) & \text{if } r_j \text{ exists and } s_i \text{ does not} \\ \frac{1}{2}(r_{j'} m_{ij'1} + s_i m_{ij'2}) & \text{if } s_i \text{ exists and } r_j \text{ does not} \\ \frac{1}{2}(r_{j'} m_{i'j'1} + s_{i'} m_{i'j'2}) & \text{otherwise.} \end{cases} \tag{3.9}$$

Hence, average scaled values are taken when reported flows are from either refereed or non-refereed countries.

To illustrate the constrained optimization framework of Poulain and Dal (2008), the generated data of Table 3.1 are used. If countries A to C are judged to be refereed nations, corrections factors for r_j and s_i where $i, j = (1, 2, 3)$ can be obtained from the partial derivatives in (3.6). These systems of equations can be expressed in matrix notation using (3.7):

$$\begin{pmatrix} 292 & 0 & 0 & 0 & -70 & -110 & -224 \\ 0 & 256 & 0 & -112 & 0 & -64 & -216 \\ 0 & 0 & 386 & -240 & -140 & 0 & -370.5 \\ 0 & -112 & -240 & 298 & 0 & 0 & -314.5 \\ -70 & 0 & -140 & 0 & 298 & 0 & -246.5 \\ -110 & -64 & 0 & 0 & 0 & 82 & -126 \\ 8 & 8 & 9.5 & 8.5 & 8.5 & 4.5 & 0 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ s_1 \\ s_2 \\ s_3 \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 56 \end{pmatrix}. \tag{3.10}$$

A function to handle migration flow data and solve these equations for any sized table and refereeing countries called `poulain` (shown in the Appendix) was programmed in S-Plus. This function requires an `array` object of data and an indicator for countries that are considered as non-refereed (`nr`). Using the `solve` command in S-Plus 6.2, the `poulain` function simultaneously determines $\boldsymbol{\theta}$ in (3.7) for all refereed countries. When provided with the data in Table 3.1, the solution in (3.10) were $\mathbf{r} = (1.0526, 1.1514, 1.0909)$ and $\mathbf{s} = (1.3114, 0.7598, 2.3108)$. These correction factors were used to scale reported counts from refereed countries using the average in (3.9). The estimated flows are shown in Stage 1a of Table 3.2. During the calculation of the values, the constraint $\sum_{i=1}^{3} \sum_{j=1}^{3} \frac{r_j m_{ij1} + s_i m_{ij2}}{2} = \sum_{i=1}^{3} \sum_{j=1}^{3} \max(m_{ij1}, m_{ij2}) = 56$ is held, where the Lagrange multiplier is a small positive value. In order to benchmark the correction factors to known definitions, as suggested by Poulain and Dal (2008), values are divided by 1.0526, that of country A's receiving data. This creates new values for $\mathbf{r} = (1.0000, 1.0938, 1.0364)$ and $\mathbf{s} = (1.2458, 0.7218, 2.1952)$, which are used to calculate the scaled averages given in Stage 1b of Table 3.2. In this table, the sum constraint of (3.9) no longer holds.

Table 3.2: Example of Stage 1 in Poulain and Dal (2008) Framework

Stage 1a: Average (Original r_j and s_i)

	A	B	C	D	E	Total
A		9.19	13.10			22.29
B	5.29		7.62			12.91
C	11.56	9.22				20.78
D						
E						
Total	16.85	18.41	20.72			56.00

Stage 1b: Average (Adjusted r_j and s_i)

	A	B	C	D	E	Total
A		8.74	12.45			21.19
B	5.03		7.24			12.27
C	10.99	8.76				19.75
D						
E						
Total	16.02	17.50	19.69			53.20

In the second and third stage, correction factors for non-refereed countries are determined from the ratio of scaled sending (receiving) data with the original receiving (sending) values of refereed countries. For example, the calculation of $r_4 = \frac{s_1 8 + s_2 8 + s_2 14}{11+8+5} = 1.9364$. A function to estimate these correction factors and estimates of flows to and from non-refereed countries, called `poulain.comp` (shown in the Appendix) was programmed in S-Plus. In order to facilitate the estimation for missing flows involving country E, data were generated using the `rpois` function in S-Plus 6.2 with rate parameter equal to 20 to reflect a (higher) stock measure. These values are given in the Stage 3 display of Table 3.3, which the `poulain.comp` function uses to estimate $r_5 = 0.6405$. In addition, sending data correction factors of $s_{j'} = (1.1266, 0.5176)$ for $j' = (4, 5)$ were estimated. Non-refereed correction factors, $r_{i'}$ and $s_{j'}$, were used to estimate the final flows using the averages of adjusted flows as in (3.9). This results in the final migration flow tables in Table 3.3.

Table 3.3: Additional Data and Final Estimates of Poulain and Dal (2008) Framework

Stage 3: Additional Data

	A	B	C	D	E	Total
A					16	
B					25	
C					18	
D					17	
E	22	22	14	18		76
Total					76	152

Final Estimates

	A	B	C	D	E	Total
A		8.74	12.45	15.63	11.98	48.80
B	5.03		7.24	10.63	10.17	33.07
C	10.99	8.76		20.21	15.64	55.61
D	12.20	11.67	16.70		13.89	54.46
E	11.69	9.52	8.81	11.44		41.46
Total	39.91	38.70	45.20	57.91	51.68	233.40

3.5 Model Component Modelling

A multiplicative component approach was applied by Raymer (2007) to estimate international migration flows between ten countries in Northern Europe in 1999. The procedure is based on modelling components as separate objects of explanation rather than the flows directly, similar to that of Willekens and Baydar (1986) or Rogers et al. (2002).

Migration flow tables can be disaggregated into four separate components (Rogers et al., 2002): an overall component representing the level of migration, an origin component representing the relative pushes from each nation, a destination component representing the relative pulls to each nation and an origin-destination component representing the connectivity between places not explained by the previous three components. Components may be derived from a log-linear regression model;

$$\log \mu_{ij} = \log \beta_1 + \log \beta_i^O + \log \beta_j^D + \log e_{ij} \qquad i \neq j, \tag{3.11}$$

where μ_{ij} is the expected migration flow from origin i to destination j. The overall effect is denoted by β_1, the origin (or row) effect by β_i^O, the destination (or column) effect by β_j^D and the interaction effect by e_{ij}. This is equivalent to the log linear model of (2.12) where $\boldsymbol{\beta} = (\log \beta_1, \log \beta_i^O, \log \beta_j^D, \log e_{ij})$ and the explanatory matrix notation containing information on flow origin and destination is implied in the constraint system of $\boldsymbol{\beta}$. The log-linear model (3.11), can be expressed in a multiplicative component form, similar to Raymer (2007),

$$\mu_{ij} = \beta_1 \beta_i^O \beta_j^D e_{ij} \qquad i \neq j, \tag{3.12}$$

where all terms have been exponentiated.

Previous implementations of multiplicative components models in migration (such as that of Rogers et al. (2002) or (Raymer et al., 2006)) have focused on the description and analysis of internal migration flows. Raymer (2007), however, uses this approach for the estimation of international migration flows through separate models for components β_i^O, β_j^D and e_{ij}. The methodology begins by estimating all origin and destination components for the migration flow table of interest using log-normal regression models and scaling reported data. This process can be separated into three stages. In the first stage, one model attempts to explain the total outflow of migrants from all nations (after scaling reported outflows totals to net migration estimates using the demographic accounting equation model). This model is used to interpolate missing marginal totals, which like origin-destination flows are often missing. In the second stage, two models attempt to explain the migration to and from all other countries in the world not included in the desired flow table, which are again used to interpolate missing values. The difference between the total flow values (a combination of scaled and interpolated estimates) from the first and second sets of models results in the total number of migrants sent and received (and hence marginal totals) for all possible flows in and out the studied countries. This allows the final estimated flow table to be expanded to include an additional row and column for flows to and from all other countries. In the third stage, the total migration within the studied flow table is derived by taking the median of the estimated total migrants sent and received, which were not previously constrained to match. This overall total is the estimate of β_1, using the total-sum reference category coding scheme recommended by Raymer (2007). Final estimates of β_i^O and β_j^D are derived by dividing marginal estimates by β_1.

In order to estimate the final model parameter e_{ij}, origin-destination migration flow data are derived by preferring receiving data over sending data where both values exist. The resulting observed values are then divided by expected values from the independence model (3.11 without the $\log(e_{ij})$ term) obtained using Iterative Proportional Fitting (IPF) algorithm of Birch (1963), where only knowledge of marginal totals and arbitrary stating cell values are required. The observed to expected ratio cannot always be calculated for every cell due to missing observed values. In order to account for incomplete data, Raymer (2007) suggests a log-normal regression model involving a dummy covariate for contiguity (for countries that share a border) to be fitted to the available ratios. This model can then be used to interpolate the missing ratio values. These ratios are then entered into the model of (3.11) in order to give a complete set of estimates for origin-destination component. In order to fit the final log-linear (Poisson) regression model of (3.11), the expected values are derived from the IPF algorithm using of β_i^O and β_j^D as marginal totals. These expected values are then regressed on the constant and dummy covariates of origin and destination with an offset of e_{ij}.

The model component framework has been applied to larger migration tables over a series of time periods by Raymer (2008) and Raymer and Abel (2008). These studies included a greater number of covariates to explain and interpolate missing values for marginal totals and flows to the rest of the world. Raymer (2008) expanded the basic origin-destination model of (3.11) by an extra dimension to estimate flows by age and sex groups. As part of the MIMOSA project, Raymer and Abel (2008) included an ad-hoc harmonization of available data as an initial step, in an attempt to account for data inconsistencies. Additional models were also built on a ad-hoc basis to aid the fit of interaction components for problematic cells. Brierley et al. (2008) conducted a study in the Bayesian paradigm with direct parallels to the multiplicative component model, using the same Northern European data as Raymer (2007). Estimates of marginal totals were fixed to the adjusted estimates found by Raymer (2007).

3.6 Discussion of Frameworks

Discussion on the presented frameworks and possible extensions is undertaken in the succeeding subsections. Comparisons between frameworks with reference to the desirable criteria outlined in Table 1.1 are outlined in the next section.

3.6.1 Constrained Optimization

The framework proposed by Poulain (1993) was the first effort to estimate an international migration flow table of comparable data. It formalized the concept that rows and columns in migration flow tables of reported data can be higher or lower due to differences in data collection and measurement techniques, and hence a correction factor can be estimated to equate data to single level. It requires some degree of expert judgement in the decision of which countries data should be included as a refereed nation. Poulain and Dal (2007) recommended that this judgement is informed by repeatedly estimating correction factors for different combinations of refereed countries. Estimates for data sources that appear unstable during this process to the analyst should be considered as non-refereed countries.

In earlier versions of the framework, the properties of estimated flows were not stated and remained unclear. Through normalizing correction factors in the first stage, Poulain and Dal (2008) were able to present a final table of estimates that possess the characteristics of the selected data sources used to normalize other values. Consequently, the constraint function of (3.4) no longer held. This may not be of deep concern for two reasons.

First, the normalized correction factors can be directly deduced without the constraint. These are estimated by reconstructing the matrix **A** and vector **b** in (3.10) to represent the set of partial derivative equations of only r_j and s_i without terms involving λ. The elements of **A** corresponding to r_1 can then be directly replaced with zero and ones in order to constrain the correction factors to unity as such,

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 256 & 0 & -112 & 0 & -64 \\ 0 & 0 & 386 & -240 & -140 & 0 \\ 0 & -112 & -240 & 298 & 0 & 0 \\ -70 & 0 & -140 & 0 & 298 & 0 \\ -110 & -64 & 0 & 0 & 0 & 82 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{3.13}$$

A function for this direct method of calculating correction factors called `poulain.direct` (shown in the Appendix) was programmed in S-Plus. It was reliant on the `poulain` function for the initial calculation of the **A** matrix. With larger flow counts, the use of substituting one to indicate fixed correction factors can create difficulties for the S-Plus function `solve`. This can be overcome by using `solve.default` which may be altered to allow the tolerance of the QR decomposition function (`qr`) used within the routine to be increased. The solution for (3.13) were $\mathbf{r} = (1.0000, 1.0938, 1.0363)$ and $\mathbf{s} = (1.2457, 0.7217, 2.1951)$, which are very similar to the parameter values from the normalized estimates for the data in Table 3.1.

Second, the right hand term in the constraint of (3.4) is justified by Poulain and Dal (2007) to allow the adjusted cell average to be equal to the maximum reported cell values. This restricts the parameter space for estimates to a level of reported migration based on a maximum measure of mixed definitions and data collection methods aggregated over all cells. This is counter to the concept of minimizing a distance function, which concentrates on the difference between (rather than totals across) reported cell values. In a similar spirit the Chi-Squared distance function of (3.3) is a weighted measure based on a denominator of observed values. These values are known to be incorrectly reported and provide an unrealistic distance measure for some flows. There exist alternative measures of distance, discussed in the next chapter, which might better capture the inequality of reported data, without a reliance on reported data.

Final estimates from the constrained optimization framework have distributions across a given row or column that are different to the original data, as flows between refereed countries are diluted to a mixture of sending and receiving patterns. This is problematic from two standpoints. First, receiving data is often believed to be of better quality (Erf, 2007; Raymer, 2007), and hence it may be more appropriate to weight the average for the calculation of the final flow estimates to reflect this consideration. The extreme of this solution would be to ignore the scaled sending data and use only the scaled receiving data for final estimates. This would result in column distributions of the original data being preserved. Second, the categorization of countries as refereed data providers is dependent on the belief that their data is consistent and complete. Taking an average of scaled data produces final flow estimates that are no longer consistent with regards to the original data distribution and hence may affect the plausibility of a final flow table estimates.

In the final stage of the methodology, correction factors are calculated and applied to a mixture of migration data. This operation is performed to obtain estimates of previously unknown cell values. As a consequence, these estimates rely on some form of migration data being available. In some cases, this might require inserting poor data solely to allow final estimates to be obtained. Taking the average of the adjusted non-refereed data assumes that the inserted data has the same row or column distributions to that of the true flows. However, non-refereed data may be known to be a poor proxy for the true flows and hence any adjustment by a scaling method may exaggerate differences in comparison to a true unknown migration flow.

3.6.2 Model Component Modelling

The multiplicative component methodology of Raymer (2007) decomposes a flow table into a number of model parameters whose values are estimated using statistical models. In the first stage, estimates of the total number of migrants sent and received are adjusted to reported flow totals derived from net migration figures. This is undertaken with the aim of creating cell estimates based on a one year timing definition. The scaling assumes that a country has the same measure in its sending and receiving data which is not always the case. For example, the Netherlands defines immigration on a six month and emigration on an annual timing criteria (Erf et al., 2006a). This can potentially inflate the receiving totals in comparison to the sending totals for which no provision is made in the framework. After the total flows have been adjusted, estimates for the total flows to nations outside the set of countries in the table are modelled in order to obtain table margin estimates for missing data. The models implemented to do this are often simplistic. In Raymer (2007) a model involving population, median age and gross national income was used to describe total outflows of ten Northern European nations. For the estimates of total flows to and from non-Northern European nations, population, gross domestic product and migration rates covariates were used. These variables are too simple when trying to replicate the framework for more dissimilar nations, and variables have no justifiable inclusion over other potential economic, geographic, demographic or social effects that may better explain total migrants leaving a nation for a particular set of nations. A more thorough method might involve the use of more complicated models to help describe the complex nature of patterns of total flows across multiple nations, as demonstrated in Raymer and Abel (2008). However, the use of multiple effects and interactions is limited by the amount of nations providing data with a total. For example, in 2006 only eight countries of the EU15 provided total counts for the number of migrants received and sent, constraining potential models to use only seven parameters to be fully identified.

Once the missing marginal estimates were obtained, all values were then scaled (again) so that overall sending and receiving totals match. These manipulations may result in final estimates not bearing much resemblance to the original reported values. Notable differences in the reported data and estimates of Northern European flows by Raymer (2007) are apparent in both the marginal totals and the distribution to and from selected countries. These differences might be justifiable for data considered unreliable, however differences were also found when comparing data that are considered to be of a good standard. For example, a comparison of reliable data sources, such as Swedish receiving data (Herm, 2006a), and the final predicted values in Raymer (2007) shows an overall drop of 2,988, a 25% fall from the reported total. The distribution of these migrants into Sweden also altered greatly, where some larger sending nations, such as Norway, were estimated to send 1,970 less migrants than reported (in line with the fall in overall total) as opposed to Finland, which was estimated to send 293 more migrants than reported. These large alterations from the distribution of known good reported data affect the plausibility of the final estimates. Raymer and Abel (2008) attempted to overcome this problem by adjusting reported flows using an iterative weighting of counts, prior to the modelling of any components. This included weights that fixed the flow values from receiving data sources believed to be of good quality.

Without flexibility in the modelling of marginal totals, interpolations for missing values may be unrealistic or even impossible. For example, Raymer (2007) had problems with overestimating the Lithuanian margin resulting in large flows to and from other nations. This over-prediction could be partly due to the simplistic models used or in the assumption that countries with missing data (such as Lithuania) all have the same relationships with the covariates used for the regression on available data. In addition, there is no restriction in place on interpolated values for migration flows to and from the rest of the world. As these values are deducted from total flows to attain a table margin, large interpolated values for the rest of world flows could potentially be greater than total flows creating a negative margin total. Once the final margins are estimated and scaled so that the overall totals match, the complete set of interaction terms are derived, again using interpolation from a simplistic model. This is based on receiving data where it exists, even if it is known to be inconsistent or reliable sending data are available.

3.7 Summary and Conclusion

The methodologies presented in this chapter take vastly different approaches to estimating a complete migration table. Comparisons of estimation frameworks are undertaken in this section using the criteria of Table 1.1. The estimates of both methodologies can be compared with respect to the completeness, consistency, reliability and measures of precision criteria. Both methods obtain estimates of complete tables, with imputations for previously unknown flows values, allowing comparisons of all flows. These are estimated using ad-hoc methods based on alternative data or existing relationships derived from simplistic models.

To account for inconsistencies, the framework of Poulain and Dal (2008) estimates correction factors for sending and receiving data, in order to scale reported counts to a definition in a selected country. Minimizing a distance function of the difference between cell values allows an explicit relationship between countries reported data to be formulated. It also allows a clear understanding of the properties of the resulting estimates. The estimates from the framework of Raymer (2007) rely on scaling total flows of available data to net migration totals suggested by the demographic accounting equation. This demands that additional row and columns for flows to all other countries outside those being studied are required, which for some nations are not always available. Once estimated, total margins for each nation are scaled, allowing the overall number of migrants to and from all countries to match. This step can radically alter the original totals of reliable reporting countries. As discussed, Sweden was one such country, which Raymer (2007) estimated to have a 25% reduction in its total flows. Although the marginal estimates match the calculated net migration, it appears unrealistic to assume that the marginal totals conform to a consistent definition due to the multiple scaling steps in the estimation of marginal totals. Inconsistent estimates might also derive from the estimation of the e_{ij} term, which is formed through a mixture of receiving and sending reported data with a variety of definitions and data collection methods.

The reliability of estimates from both frameworks can be tentatively compared with values and distributions of good quality reported data. Poulain and Dal (2007) estimated migration flows between 28 European nations in 2004, (for the paper in 2008 no estimates were given). For countries that were considered to have good quality data, such as Sweden, estimates tended to be larger than receiving and sending reported data. These differences are constant (and in most cases smaller) for flows to and from non-refereed countries due to estimated correction factors close to unity. For flows between non-refereed countries estimates tended to be further from the reported (stock) data and on some occasions are reliant on very large correction factors. Consequently, unreliable estimates for originally missing data can appear when applied to different data. As discussed previously, the estimates from the framework of Raymer (2007) appeared unreliable when distributions are compared with good quality data. A further measure of reliability could be undertaken by comparing fully estimated migration flows tables across time. Correction factors for refereed countries by Poulain and Dal (2008) are estimated in the succeeding chapter, and demonstrate some stability across time. Neither frameworks considered in this chapter allow estimates of precision measures to be obtained.

The methodology in both frameworks can be compared with respect to the use of model based imputation methods, allowance for expert opinion, replicability for other users and flexibility to alternative data from different countries and time periods. The constrained optimization procedure imputes missing data in an ad-hoc manner, relying on alternative data to be scaled according to correction factor estimates. In the multiplicative component framework, missing components are interpolated from simple model based on available data. Consequently, it is assumed that they share the same relationship as the observed data. More considered methods exist in statistics, such as the Expectation-Maximization algorithm of Dempster et al. (1977), which can more fully account for incomplete data. This algorithm can be applied in either methodology to estimate missing data based on models for the scaled flow data from refereed countries or for components of a migration table. The former of these will be further studied in Chapter 5 of this thesis. An alternative statistical approach to handle missing data could be undertaken in the Bayesian paradigm.

Brierley et al. (2008) conducted such a study with direct parallels to the multiplicative component model, using the same original data and marginal totals (which were fixed) as Raymer (2007). Discussion of this method is left to the concluding chapter of this thesis, as a concentration on harmonizing and modelling incomplete data in a Frequentist approach in the remainder of this thesis is taken.

The constrained optimization framework allows a small degree of expert opinion in the selection of refereed countries. Poulain (1999) selected refereed countries from an analysis of the stability of correction factor estimates when systematically excluding data sources. No explicit level of stability is mentioned and hence some level of expert opinion can be used to determine which countries can be used to calculate refereed correction factors. This feature is expanded further in the next chapter, to harmonize data that is reported to be of good quality in recent literature on international migration flow statistics. The approach of Raymer (2007) allows expert opinion to be used in the selection of covariates in models for the interpolation of components for missing data. Brierley et al. (2008) demonstrated that in a Bayesian framework, expert opinion can be fully incorporated to alter estimates that are believed to be from unreliable data sources.

The replicability of the methodology of Poulain and Dal (2008) is considerably better than that of model component modelling. As previously mentioned a S-Plus function was created to quickly estimate correction factors and estimated flow tables. The methodology of Raymer (2007) was more complicated, with multiple stages of data manipulations, interpolation and model fitting. These can cause errors in the implementation of the framework leading to different estimates. Unlike the constrained optimization framework, the models to interpolate missing margins and interaction components require extra covariate information, adding further complication. The Raymer and Abel (2008) extension of this framework adds further stages, some of which are ad-hoc and dependent on estimates from previous stages.

Both frameworks discussed in this chapter concentrated on European data. The framework of Poulain and Dal (2008) had been previously applied (with alternative distance functions and constraints) to alternative migration tables of different sizes and in time periods. Due to the ad-hoc nature of estimating missing cell values it is dependent on the availability other sources of migration data, such as stocks. These may not always be available and up to date, which could severely affect the reliability of final estimates. The model component framework has also been applied to larger European migration tables over a series of time periods by Raymer (2008) and Raymer and Abel (2008). Both of these studies included a greater number of covariates (allowed by working with a larger table) to explain and interpolate missing values for marginal totals and flows to the rest of the world. Additional models were also built on a ad-hoc basis to aid the fit of interaction components for problematic cells.

In conclusion, two very different frameworks exist to estimate international migration flow tables. Both fail to satisfactorily address all the criteria for migration flow table estimation methodologies set out in Table 1.1. However, some elements of the methodologies provide useful guidance that could be used in a more comprehensive framework. Most notable was the underlying concept introduced by Poulain (1993) of estimating correction factors to adjust reported data to a consistent level. This was based on an underlying assumption that differences in the reported counts can be considered as non-random measures of the discordance in the collection and measurement of migration flows between reliable data sources. The following chapter will explore this aspect further, investigating different distance measures, the use of alternative constraints and using current research into data sources to further improve the comparability of international migration flow data.

Chapter 4

Overcoming Inconsistencies in International Migration Flow Tables

4.1 Introduction

The lack of comparability in international migration data can be traced to the multi-dimensional nature of migration (Goldstein, 1976). As a result, national statistics institutes have developed measures of migration solely suitable to their domestic priorities. When international migration data is compared across multiple countries in a single time period, inconsistencies in reported flow values between data sources are apparent. Constrained optimization studies such as Poulain (1993) attempt to harmonize international flow data by estimating correction factors to adjust reported data to a consistent level, where differences in the reported counts are considered as non-random measures of the discordance in the collection and measurement of migration by national statistics institutes.

The analysis and application of constrained optimization methods for international migration flow data has been predominantly limited to a single time period offering only a loose guidance on its application and neglecting the underlying causes of the incompatibility in international migration flow data. The application of alternative distance and constraint functions has remained partly ignored, driven by concerns of estimating missing values for international migration flow tables of comparable data. This chapter concentrates solely on inconsistent data issues to develop a methodology for the estimation of consistent migration flow data that are comparable across multiple nations. Included in the study is an exploration of alternative distance and constraint functions. These will be analyzed for reported flows over a series of time periods to allow added information to inform the estimation of correction factors.

This chapter commences by presenting a series of migration flow tables for comparisons over time for 15 countries of the EU before the expansion of May 2004 (EU15). Next, expert opinion on the characteristics of data from these countries and levels of counts of migrations with unknown origins or destinations are presented. These allow a clearer explanation of the differences in sending and receiving data and help in determining where reported counts can be scaled to a comparable level. For some reported data, expert opinion indicates that counts are inaccurate and hence a scaling of their values is not considered as it would accentuate errors already present. The following section introduces a methodology for creating comparable estimates from reliable data sources using constrained optimization routines in statistical software. These allow the estimation of correction factors to be easily obtained and with a great deal of flexibility in the specification of distance and constraint functions. This methodology is applied to a series of international migration flow tables to estimate alternative sets of correction factors for EU15 nations in two stages. First, correction factors for different constraint sets are estimated using the same distance measures. Second, correction factors for a range of distance measures on the same set of constraints are estimated. Comparing estimates across time, the robustness of harmonization methods to changes in migration flows are assessed, under the assumption that the sources of inconsistencies have remained the same. In the final section, a generalization of the distance measures and constraint sets over time is carried out to enable a larger number of observations to be used in calculating final estimates of correction factors. This yields an incomplete set of international migration flow tables which will facilitate statistical modelling in following chapters, allowing imputations for missing flow values to be obtained for a complete harmonized table.

4.2 International Migration Flow Data for the EU15

International migration flow data may be obtained from a number of international organizations. One of the most comprehensive collections is provided by Eurostat. Data are collected from individual national statistics 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 data base, New Cronos which is available online. The reported counts of these flows can also be found in publications of individual national statistics institutes, the CoE and Système d'Observation Permanente des MIgrations (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 data bases. The cause of this difference is not known due to insufficient documentation (Kupiszewska and Nowok, 2008).

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

Reported values can be represented in two separate migration tables similar to Table 3.1 or as double entry tables shown in UN (1976), Kelly (1987) or Nowok et al. (2006). For the 2006 data, a double entry matrix is displayed in Table 4.1, where origins are shown on the vertical axis and destinations on the horizontal axis. Countries are labelled according to three-letter classification by the International Standardization Organization (ISO). Each cell comprises of two counts when both are available. The top values are collected by the receiving destination countries and hence are read vertically. In the same manner, the bottom value in each cell contains the origin reported values, as collected by the sending nations, forming a horizontal pattern. As noted in the previous chapter, reported counts may be very similar, such as the flow from Austria to the Netherlands, or very different, such as the flow from Austria to Germany. These altering differences give a confusing impression as to which data source, if any, to be preferred.

When data is collected over time a graphical representation of cell values allows an easier viewing of migration levels and data issues. Plots of a series of migration flow tables for the EU15 are shown on a logarithmic scale in Figure 4.1. Red lines represent receiving country data and blue lines sending country data. Origins are shown on the vertical axis and destinations on the horizontal axis.

When compared over time it is evident that some nations such as Belgium or France never provide receiving data, and hence no red line appears in their columns. Other nations such as Greece or Portugal never provide sending data, and hence no blue line appears in their rows. Ireland consistently provides data only to Great Britain, with exception of the last time period. In origin-destination cells where both sets of data are reported, the lines are fairly parallel, a feature illustrated in Kupiszewska and Nowok (2008) for selected flows between nations with good quality data collection procedures. Non-parallel lines are visible for reported flows to and from some nations such as Great Britain, where British counts tend to be more volatile than their reporting partners. For larger flows, such as German and Spanish flows to and from Great Britain, British data are more volatile when plotted on a non-logarithmic scale.

Reported flow values tend to be highest to and from of countries with the largest populations such as Germany, Great Britain, France, Italy and Spain. Values between neighbouring countries, such as Netherlands and Belgium or Germany and Austria, tend to be larger than other values in the same row or column. Of the 1050 cells (corresponding to a 15 × 15 non-diagonal mobility table over 5 years), 870 had values from at least one reporting partner. In 332 cells, data from both sending and receiving countries were available for which none reported the same value. In 225 cells there were no reported values from either country. For 20 origin-destination combinations (out of a possible 210) there is no data reported in any year.

A plot of the counts when both nations report data is shown in Figure 4.2. On the right hand panel are counts as given in Figure 4.1, whilst on the left hand panel the counts

Table 4.1: Reported Double Entry Migration Flows from each Origin-Destination Combination of the EU15 in 2006

Receiving values in top half of cells, Sending values in bottom half of cells. Countries labeled according to three-letter classification by the ISO (2006)

	AUT	BEL	DNK	FIN	FRA	DEU	GRC	IRL	ITA	LUX	NLD	PRT	ESP	SWE	GBR
AUT			344	108		14719				9	591		964	324	
		257	186	284	718	9244	403	155	1195	46	599	255	638	430	1086
BEL	269		529	204		4115				605	6149		3391	429	
DNK	189			353		2563				17	488		1076	6432	
	238	413		362	1145	2690	177	256	675	166	615	129	1593	6413	3538
FIN	303		456			1984				3	336		770	3092	
	87	187	429		334	711	47	240	199	90	218	24	528	3071	1190
FRA	1038		1755	408		19095				927	3357		15604	1216	21821
DEU	18467		4471	981						455	10424		18417	4090	20505
	18604	4540	3115	2146	17790		15653	2330	26807	1864	9189	7014	16734	3934	17319
GRC	368		252	80		8957				4	1175		694	640	
IRL	147		270	132		1724				14	471		1972	310	
ITA	1683		1044	290		20130				58	1966		12625	789	
LUX	78		153	53		2611					170		158	84	0
NLD	862		939	272		14054				29			5950	1319	
	675	12008	599	363	3842	11006	649	713	1422	234		982	4418	1194	9032
PRT	319		252	67		5640				517	1696		19304	211	
ESP	758		1845	702		14219				25	3372			1459	14736
	201	1101	181	106	2901	2732	89	521	1201	99	920	1289		209	3509
SWE	608		3629	3448		3181				11	787		2102		
	437	402	3456	3365	1126	1792	643	315	544	159	598	140	1499		4479
GBR	1285		3235	1000		12903				35	5552		41876	3295	
					25962								35277		

Figure 4.1: Reported Migration Counts (000's) for each available Origin-Destination Combination of EU15, 2002-2006 of the Logarithmic Axis.

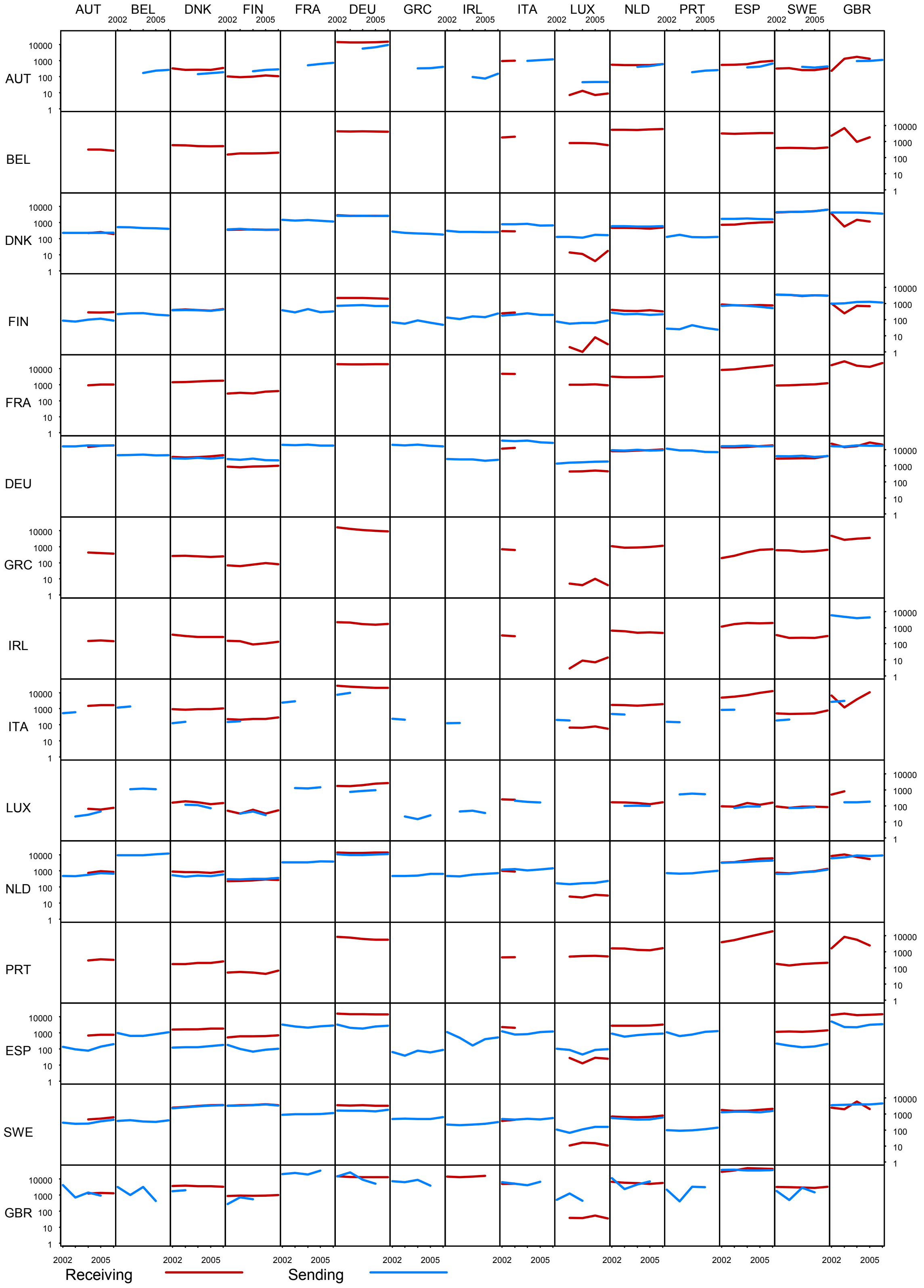
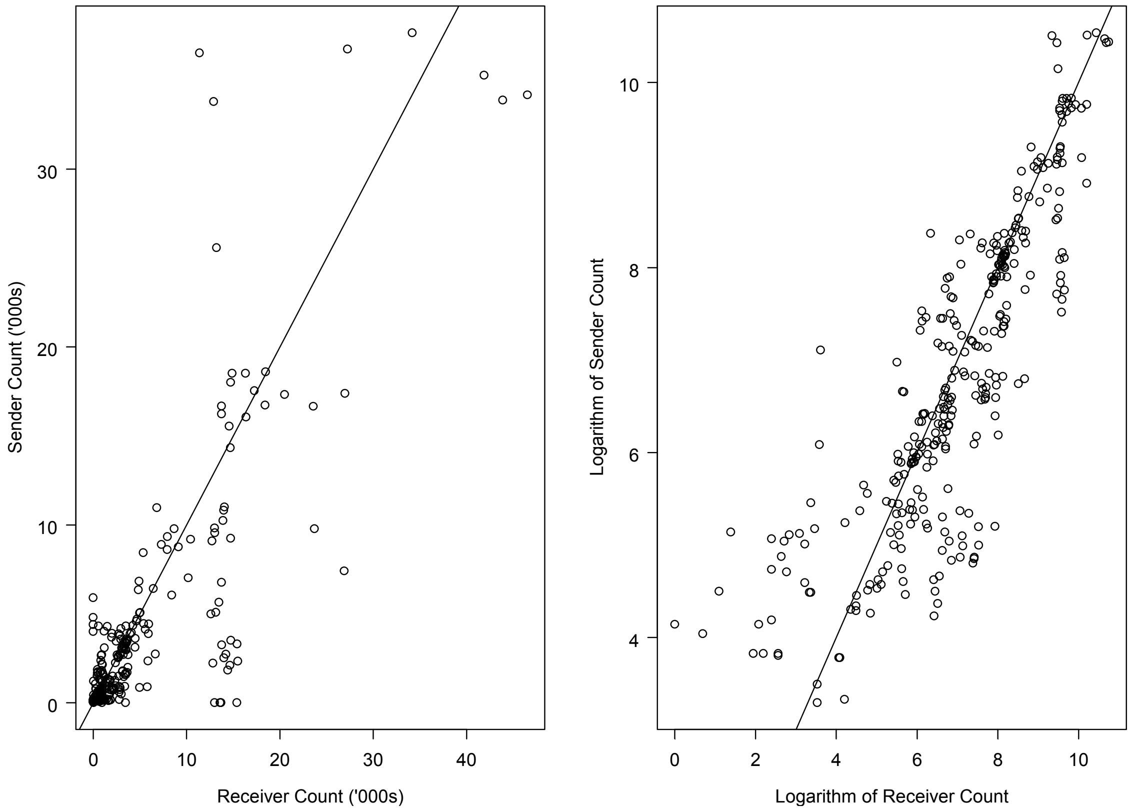


Figure 4.2: Sending and Receiving Migration Flow Counts (left) and Logarithm of Counts (right) for EU15 Countries, 2002-2006.



in thousands are displayed. These plots demonstrate the correlation between sending and receiving data when both are reported. If all data sources produced comparable data (from identical data collection system and measurement) all points would lie on the solid diagonal line where sender and reported values are equal. The logarithmic plot demonstrates that the distance from this equality is not necessarily influenced by the size of the reported flows when a transformation of the reported data is taken.

Some of the smallest differences occur for flows between the Nordic nations of Sweden, Finland and Denmark. These nations all use registration systems to collect migration data. An exchange system is in place for the reporting of movements between Nordic countries, as migrants are only registered in one country at a time (Herm, 2006b). Consequently, data for the number of migrants sent from one of these nations is recorded by the country of destination, rather than origin. Reported counts of migrants sent between Nordic countries, as collected by the sending data source, are unavailable. Differences in these counts are attributed to dual citizens and time delays for migrations occurring at the end of the year (Nowok et al., 2006).

4.2.1 Ratings of Migration Data for EU15

In order to obtain a comparison of the European migration flow data, Erf (2007) provided subjective judgements by three characteristics: definitions of migration, measurement

Table 4.2: 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

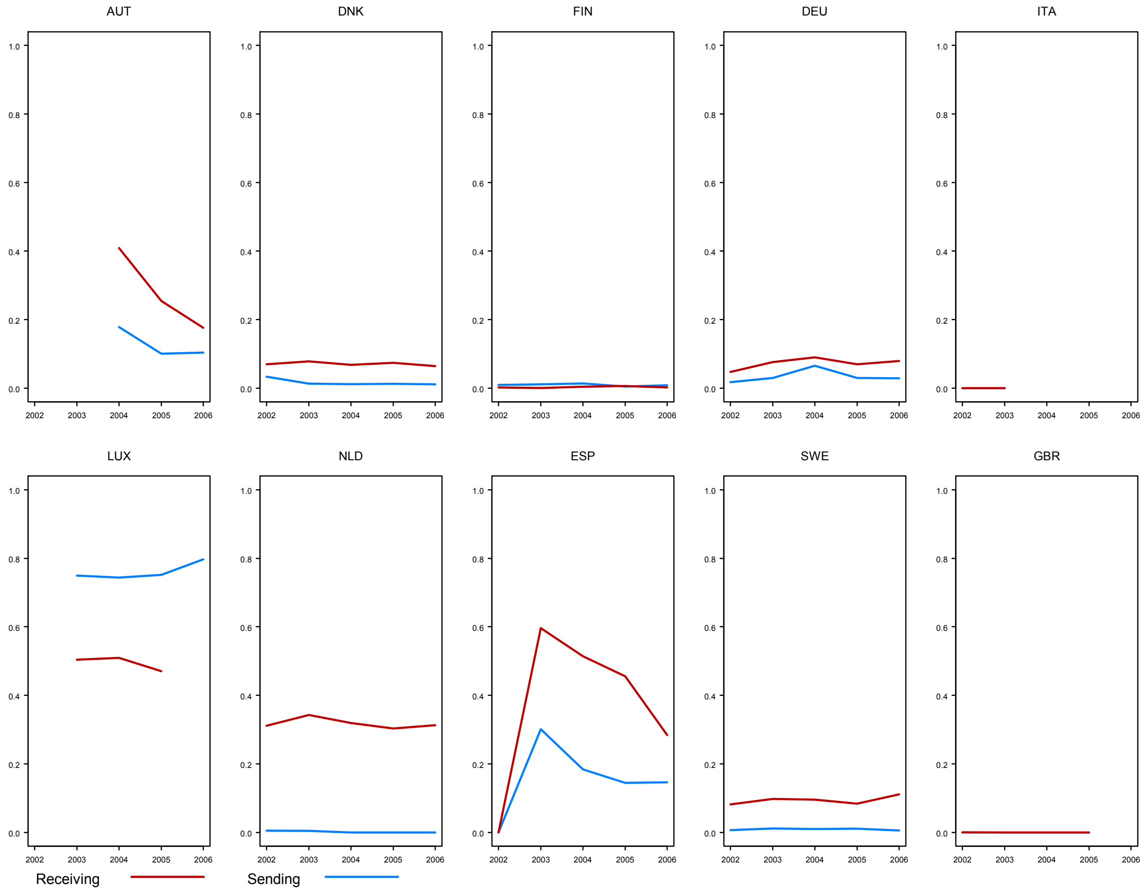
Scores in parentheses are for non-national, when national and non-national data are collected differently.

systems and intended coverage. For member nations of the EU15, ratings for both receiving and sending data between 2001 and 2006 are shown in Table 4.2. 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 believed to be present in the measurement systems used. Scores for accuracy are based on the coverage of the target population and the collection, production 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 the completeness and accuracy of receiving and sending data were the same. Greece fails to provide any receiving flow data and both France and Greece do not publish any sending migration data throughout the time period. For 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 period.

4.2.2 Data Dissemination in the EU15

Plots of the available counts of migrants with unknown origins or destinations. as a proportion of total sending and receiving countries, are shown in Figure 4.3 for EU15 nations between 2002 and 2006. Totals for this calculation were given by the New Cronos

Figure 4.3: Proportion of Migrant Origins or Destinations Unknown for Available Receiving and Sending Data of EU15 in 2002-2006



data base, which correspond to totals of all flows (including the counts of migrants with unknown origins or destinations). As with the flow data, unknown counts are reported according to local definitions and data collection methods. With the exception of Luxembourg, these plots demonstrate that sending data tend to have a lower proportion of unknown destinations in comparison with the unknown origins in receiving data. 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 unknowns are created from the non-reporting of departures by emigrants and the non-collection of country of origin by local municipalities (from which national level data is aggregated Perrin and Poulain (2006a)). For Spanish data, there is a notable change in the level of unknowns between 2002 and 2003, with an increase from 69 (and 6) to 202,256 (and 38,339) received migrants (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 and Thierry, 2006b). In the Netherlands, emigrants have to deregister from their municipal data base 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 counts (Erf et al., 2006a).

4.3 Methodology for Creating Comparable Data from Reliable Data Sources

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

(a) correction for unknown counts,

(b) harmonization of reliable data,

Each stage is outlined in turn.

4.3.1 Counts of Unknown Migrant Origins and Destinations

As previously discussed, international migration flow data are accompanied by a count of migrants with unknown origins or destinations. If we let migration flow tables of such data be represented by array n_{ijk}, where i indicates migrant origin, j indicates migrant destination and $k = 1, 2$ indicates receiving and sending flow tables respectively. For the receiving flow table there exists a row n_{uj1} which contains the counts of unknown flows collected in destination j. In the same respect, the sending flow table there exists a n_{iu2} which contains the counts of unknown flows collected in origin i. In order to account for these unknowns and thus avoiding bias towards data sources with no unknowns, corrected migration flows, m_{ijt} can be derived as follows,

$$\begin{aligned} m_{ij1} &= n_{ij1} + \left(\frac{n_{ij1} n_{uj1}}{n_{i+1} - n_{uj1}} \right), \\ m_{ij2} &= n_{ij2} + \left(\frac{n_{ij2} n_{iu2}}{n_{+j2} - n_{iu2}} \right), \end{aligned} \tag{4.1}$$

where the index $i, j = +$ denotes total flows including unknowns counts. This allocation assumes that unknown counts are missing at random among all international origins or destinations. If a certain type of migration, such as inter-continental moves, are more likely to be captured and reported in the data collection dissemination then this allocation would discriminate against more local moves whose origin or destination may not be known.

4.3.2 Constrained Optimization

Differences in counts between nations with better quality data can be considered as fixed, where data production techniques do change over time. Thus, a distance measure of these differences represents the non-random discordance in the collection and measurement of migration flows between any two national statistics institutes. Poulain (1993) took a similar view in his attempts to harmonize migration data, whereby all reliable data were considered to be influenced by some data source specific correction factor. As outlined in the previous chapter, correction factors can be estimated to minimize these distances using a constrained optimization method. Correction factors can then be used to scale reported counts to a comparable level.

In this chapter, the constrained optimization method is extended to alternative distance measures, constraint sets and generalized across a series of migration tables. This is implemented in five stages

(a) select reliable data and constraints on the basis of expert opinion,

(b) estimate correction factors for different distance measures and time periods,

(c) select a distance measure associated with the set of correction factors that are most stable over time,

(d) generalize the distance measure to estimate a single correction factor for each data source over the entire period,

(e) use the correction factors to scale reported data.

Each stage is discussed in turn.

Data sources for which distances can be considered as fixed are selected using expert opinion. In this chapter, the rankings by characteristics outlined in Table 4.2 are used to select data sources that provide reliable reported counts. When data sources are considered insufficient or data are not available, reported counts are ignored. This selection criteria provides a set of migration tables (that may be non-square) of reliable sending and receiving migration flow data. Expert opinion is also used to select a correction factor(s) for at least one of the reliable data sources which will be constrained to equal one. This allows other correction factors to be interpreted as the effect of different measurement and collection methods in each data source with reference to the constrained data source(s).

Estimates of non-constrained correction factors are determined using constrained optimization routines in statistical software. These allow a great deal of flexibility in the estimation of correction factors for a range of distance measures and constraint sets. In this chapter, the `nlminb` function in S-Plus 6.2 is used. This routine can find a local minimum for a twice differentiable function within a multi-dimensional bounded parameter space. Required arguments for the procedure are the function to be minimized and suitable starting values for the parameters. If unrealistic starting values are used or the function is complex, the solution may not be correctly determined. Gradient and Hessian functions may also be considered by the routine to obtain solutions quicker. When derivative functions are not available, the `nlminb` routine implements a quasi-Newton optimizer to find parameter values such that the given function is minimized. Alternative routines such as `fmincon` in Matlab also employ a quasi-Newton optimizer for constrained minimizations with multiple parameters.

A quasi-Newton optimizer operates in a similar manner to the Newton optimizer of (2.42), discussed in Chapter 2. When the gradient $\mathbf{v}$, and Hessian, $\mathbf{H}$, are not known, initial values are calculated numerically. Approximations of $\mathbf{v}$ are taken using finite differencing. Algorithms also exist to approximate the Hessian matrix and its inverses when unknown, such as the BHHH routine of Berndt (1974) (see Nocedal and Wright (1999, p194-210) or Skrondal and Rabe-Hesketh (2004, p181-2) for more details). These algorithms allow the quasi-Newton optimizer to modify approximations of $\mathbf{H}$ in each iteration by combining the most recently observed $\mathbf{v}$ and $\mathbf{H}$ with existing knowledge embedded in the current Hessian approximation.

Correction factors estimated from alternative methods, such as the `poulain` and `poulain.direct` functions discussed in the last chapter, can be compared with the `nlminb` function. In this chapter, this comparison is undertaken by minimizing the Chi-Squared distance function of (3.3) using the `nlminb` routine. In addition, different constraint sets and the effect of ignoring data exchanges, such as those between Nordic countries, are analyzed. These can provide a further insight on the effect of multiple constraints and optimization procedures on the estimates of correction factors.

Comparisons between sets of estimated correction factors can be drawn from several sources. Plots of correction factors over time allow a clear illustration of the effect of different distance functions, estimation methods and constraints. As differences between reliable reported flow data are considered fixed over time, the most effective distance measure should provide the same correction factors in each year. This stability can be empirically summarized by considering each set of correction factors $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^T = (\log(\mathbf{r}_t), (\log(\mathbf{s}_t))^T$ for time period t. The variance within correction factors over time can thus be estimated as,

$$\frac{\sum_{d=1}^{p} \sum_t (\boldsymbol{\theta}_{dt} - \bar{\boldsymbol{\theta}}_d)^2}{n-p}, \tag{4.2}$$

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 (4.2). This allows the variation between larger correction factors to have an equal effect as smaller correction factors.

As definitions and collection methods of all the reported data used in the estimation are assumed fixed over time, the distance measure that possesses the smallest variation can be regarded as the best measure for a constrained optimization of migration flow data. For such a measure, a set of single correction factors for each data source over an entire series of tables can be estimated. To estimate these correction factors, consider a series of double entry migration flow tables noted as m_{ijtk}, where i indicates referee origin countries, j indicates referee destination countries, t the time period and $k = 1, 2$ indicates receiving and sending flow tables respectively. Each column of the receiving data table, m_{ijt1} can be assumed to be influenced by some correction factor r_j that scales the value of reported counts based on the collection methods and definitions used in the respective data sources. In the same respect, the sending table m_{ijt2} is influenced by row factors s_i.

Estimates of these correction factors can be derived from a constrained optimization on the selected distance measures generalized over the entire array of reliable reported data.

Final correction factors are used to scale reported data as such,

$$y_{ijt} = \begin{cases} r_j m_{ijt1} & \text{if } r_j \text{ and } m_{ijt1} \text{ exist at time } t, \\ s_i m_{ijt2} & \text{if } s_i \text{ and } m_{ijt2} \text{ exist at time } t \text{ and } r_j \text{ does not,} \\ z_{ijt} & \text{otherwise,} \end{cases} \tag{4.3}$$

where z_{ijt} represents a subset of y_{ijt} that have missing values depending on the lack of corresponding correction factors. The application of correction factors in (4.3) is an alternative strategy to the approach suggested by Poulain (1999) who took an average of the scaled data. The correction of receiving data, when sending data are available, will result 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; 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. Consequently, an altered distribution of flows will be estimated across a row when compared with the original data. This alteration will be to greater effect than under an averaging of corrected flows, but will provide estimates for counts in destinations where no reliable receiving data are available.

4.4 Estimating Comparable Data from Reliable Data Sources

In order to estimate comparable data from reliable data sources, reported counts are adjusted for unknowns produced in the dissemination of data by national statistics institutes. Non-linear optimization routines are then applied to the EU15 data. This is undertaken in two stages. First, different constraint sets and estimation methods are tested using the same distance measures. This provides a better understanding of the effect of the `nlminb` function in comparison with other constrained optimization techniques presented in the previous chapter. Second, a range of distance measures on the same set of constraints (suggested by data rankings) are estimated. Comparing results over time allows the robustness of distance measures to changes in migration flows to be determined from the within variance statistic of (4.2).

4.4.1 Correction for Unknown Counts

All unknown counts, displayed in Figure 4.3 are distributed to origins and destinations using the equations in (4.1). This reduces the difference between some reported counts, such as flows into Luxembourg, where reported receiving data are persistently lower than sending data of corresponding origin countries. For Spanish data, the addition of greater unknown counts in years previous to 2002 increased counts to similar levels as the 2002 counts.

4.4.2 Comparison of Estimation Methods and Constraint Sets

The analysis of methods to estimate correction factors to minimize the Chi-Squared distance function was undertaken for reliable EU15 migration flow data in Figure 4.1. This included data from all sources ranked with scores of at least reasonable, for completeness and accuracy characteristics, in Table 4.2. Since not all data, from sources considered reasonable, were available for all time periods the size of tables and consequently the number of estimated correction factors, changed for each time period. To compare estimation methods three sets of correction factors were estimated using

(a) A total constraint and normalization to Swedish receiving data proposed by Poulain and Dal (2008) using the `poulain` function,

(b) A single constraint to Swedish receiving data using the `poulain.direct` function presented in the previous chapter,

(c) A single constraint on Swedish receiving data estimated using the `nlminb` function.

To compare constraint sets and the effect of ignoring data exchanges, estimates from (c) can be compared with correction factors estimated using

(d) Multiple constraints on correction factors corresponding to data sources ranked with scores of good for timing, completeness and accuracy by Erf (2007) estimated using the `nlminb` function,

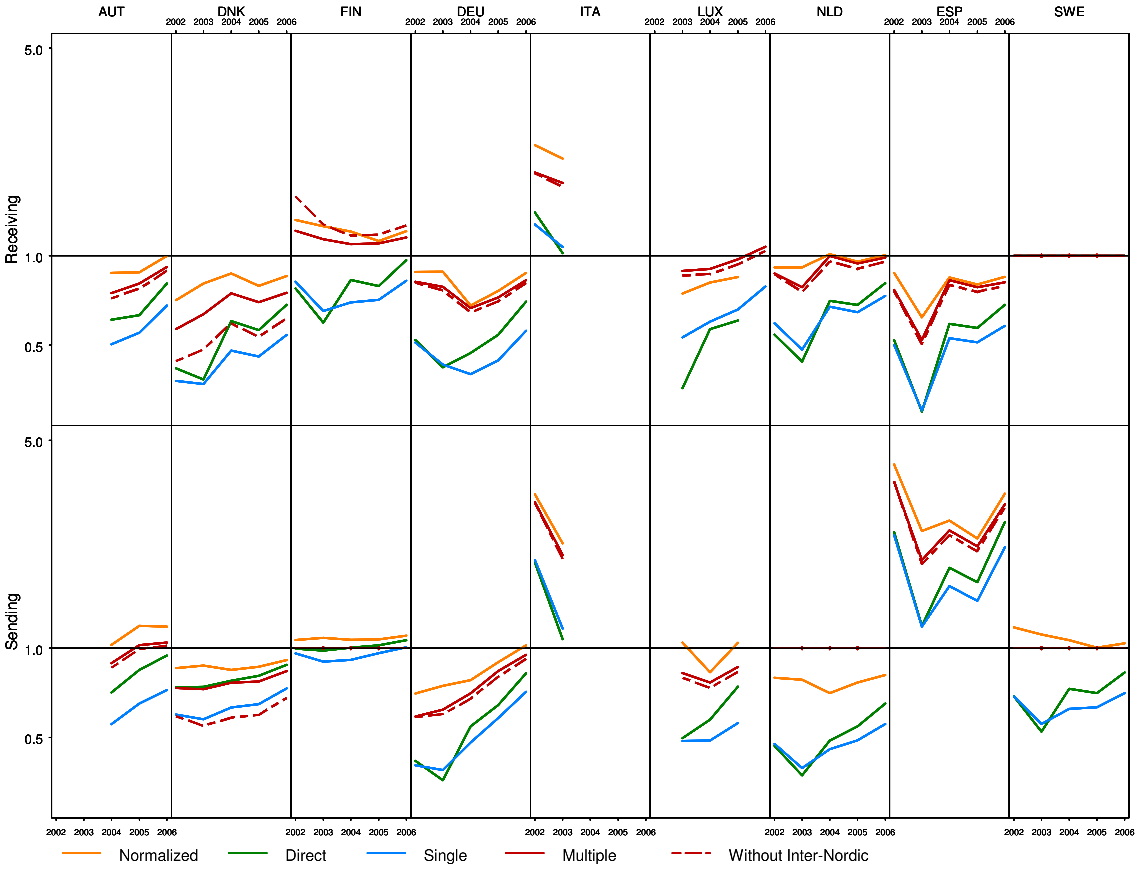
(e) A repeat of (d), excluding data for flows between Nordic countries.

For the last three applications (all of which use the `nlminb` function) lower and upper bounds were defined for all parameters to be between 0.1 and 10 with the exception of correction factors with constraints where both bounds were set to 1.0. All initial parameter estimates for the function were set to 1.0. The S-Plus/R Chi-Squared distance function is shown in the Appendix.

Correction factors from the different estimation methods were obtained in each time period. In all cases the function successfully converged to a minimal distance value. Comparisons of estimated values are displayed in Figure 4.4. These plots illustrated some clear differences in correction factors resulting from different estimation techniques.

Estimated correction factors for Swedish receiving data from all estimation methods are unity and hence their plots overlap. Estimates from the normalization to Swedish receiving data proposed by Poulain and Dal (2008) (calculated using the `poulain` function) are comparatively higher than all other methods illustrated. This is caused by the constraint on the summation of reported values which inflates correction factors to levels artificially high to meet the total constraint, before the normalization to Swedish receiving data is taken. Consequently, the minimal Chi-Squared distance in each time period is often higher than alternative methods, as shown in Table 4.3. Estimates from the `poulain.direct` function (shown by the green line of Figure 4.4), is similar to estimates from the `nlminb` function (shown by the blue line). In the latter, the within data source variance (provided

Figure 4.4: Receiving (r_j) and Sending (s_i) Correction Factors from 2002-2006 using the Chi-Squared Distance Function



in Table 4.3) is lower whilst correction factors tend to be higher. These differences are the result of two features. First, the direct method relies upon the `solve` function of S-Plus, and hence require square matrices during the inversion process, unlike the non-linear minimization function. As a result, this method is unable to provide correction factor estimates for Luxembourg's sending data in 2006 as no receiving data is reported. This results in less distance measures considered in the `poulain.direct` function. Second, the quasi-Newton method considers an estimate of the second differential of the distance function in the estimation of correction factors, unlike the method of Lagrange Multipliers. This allows estimates of correction factors to fully consider the curvature of distance function when searching for minimal values.

In order to compare different constraint sets, multiple correction factors for Swedish, Finnish and Dutch sending data (as well as Swedish receiving data), were all fixed to 1.0 in the `nlminb` routine, as all data sources were given ratings of good for timing, completeness and accuracy by Erf (2007) (Table 4.2). The multiple constraints lead to a reduction in the variance and higher minimum distances in comparison to the correction factors estimated with a single constraint on Swedish receiving data. Higher minimum values are caused

Table 4.3: Summary of Constrained Optimization Methods on the Chi-Squared Distance Function

Constraints	Estimation	Distance at Minimum					Variance
		2002	2003	2004	2005	2006	
Single	`poulain`	3754	5122	3236	3252	2554	0.0850
Single	`poulain.direct`	2240	2754	2141	2239	1951	0.1376
Single	`nlminb`	2016	2468	1887	1965	2075	0.1154
Multiple	`nlminb`	3827	5095	3706	3596	3489	0.0824
Multiple*	`nlminb`	3167	4494	3282	3101	3087	0.0868

*Excluding Inter-Nordic Flows

by correction factors for Dutch and Swedish sending data becoming constrained to unity, where previously their values were below one.

The removal of distance measures, derived from inter-Nordic data leads to a small increase of the variance in correction factors. Correction factors with and without these measures were very similar with the exception of Danish and Finnish (receiving) data for which correction factors were estimated further from unity when inter-nordic flows where ignored. Minimal distance measures (in Table 4.3) with multiple constraints were lower in most years when inter-Nordic flows were dropped. This is due to fewer observed measures considered in the distance function.

4.4.3 Comparison of Distance Measures

Alternative distance functions, to the Chi-Squared distance measure, could provide more stable correction factors over time, and hence better reflect the assumption that data collection methods and definitions remain constant. The range of distance functions considered ($f(r_j, s_i|m_{ijk})$) for the routine are shown Table 4.4.

Table 4.4: Alternative Distance Metrics and Estimated Variance from 2002-2006 Data

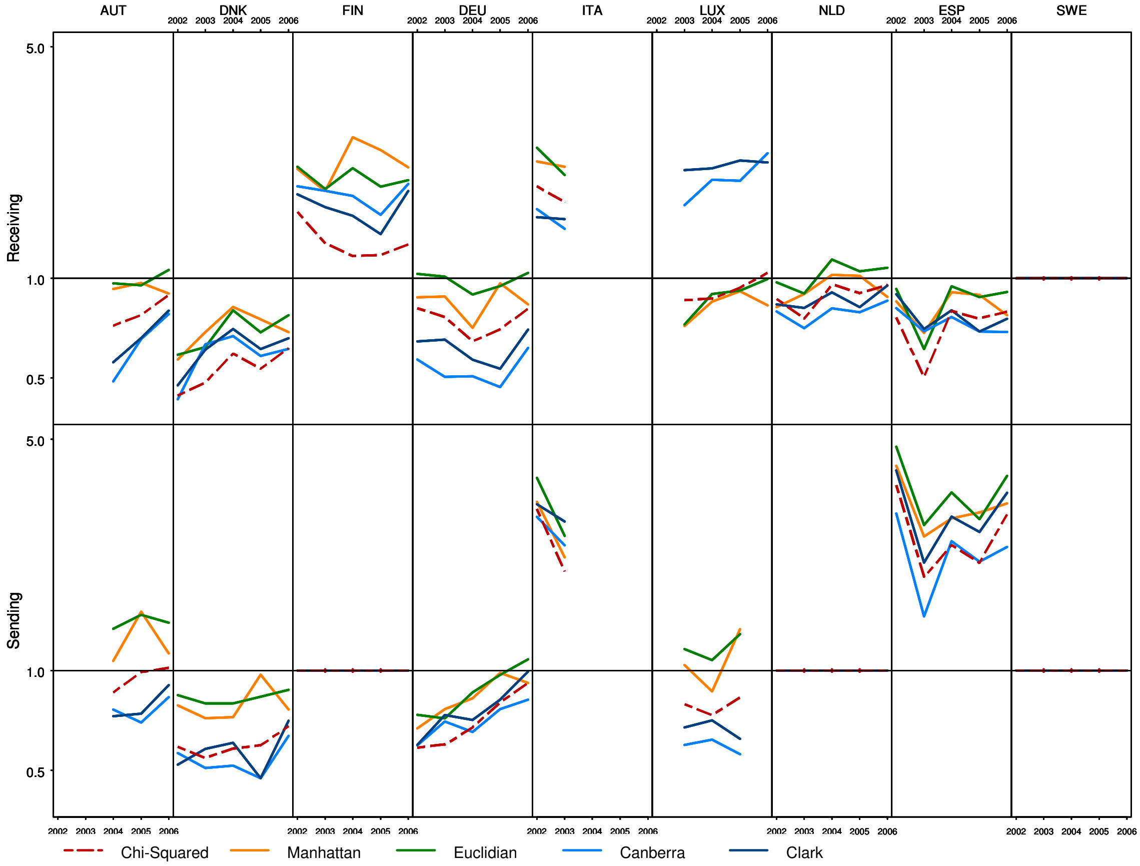
Distance	$f(r_j, s_i\|m_{ijk})$	Variance
Manhattan	$\sum_{i,j} \|r_j m_{ij1} - s_i m_{ij2}\|$	0.0877
Euclidean	$(\sum_{i,j} \|r_j m_{ij1} - s_i m_{ij2}\|^2)^{\frac{1}{2}}$	0.0944
Canberra	$\sum_{i,j} \frac{\|r_j m_{ij1} - s_i m_{ij2}\|}{r_j m_{ij1} + s_i m_{ij2}}$	0.0740
Clark	$\sum_{i,j} \frac{\|r_j m_{ij1} - s_i m_{ij2}\|^2}{(r_j m_{ij1} + s_i m_{ij2})^2}$	0.0892

The first two measures considered were the Manhattan and Euclidean measures, (the latter equivalent to the Euclidean distance of (3.2) used by Poulain (1993)). The general form of these measures are also known as Minkowski distance of order p or p-norm distance,

$$(\sum_{i=1}^{n} |r_j m_{ij1} - s_i m_{ij2}|^p)^{1/p} \tag{4.4}$$

where $p = 1$ or $p = 2$ for a Manhattan and Euclidean distances respectively (Deza and Deza, 2006, p126). Both provide equal weighting for each reported flow, and hence an optimization procedure depends solely on minimizing all distances regardless of the flow sizes. The third and fourth distance functions are based on the Canberra and Clark measures (Lance and Williams, 1967). These use weightings to allow differences to be measured relative to the scaled reported data.

Figure 4.5: Receiving (r_j) and Sending (s_i) Correction Factors, 2002-2006 for Different Distance Functions



Estimates for the correction factors from these measures in each time period are given in Figure 4.5. As with the final estimates of correction factors in the previous subsection, these were estimated on tables of reliable flows (adjusted for unknown counts) between 2002 and 2006, ignoring inter-Nordic data and using the `nlminb` routine. In all cases the function successfully converged to a minimal distance value.

For the first two measures (orange and green lines respectively), estimates tend 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 (light and navy blue) also resemble each other and on occasions differ from 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 tend 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 not alter the direction of scaling. Spanish estimates fluctuate greatly in comparison with others, with values for most distance measures falling for 2003. This might be related to the changes in the level of unknowns discussed earlier.

For comparative purposes, the correction factors from the Chi-Squared distance function in Figure 4.4 are also plotted using the dashed red line. The estimates from this distance measure are regularly between the weighted and non-weighted versions, as the denominator is the summation of unweighed flows. Its variance, shown in Table 4.3 is similar to estimates from the Manhattan distance function. The smallest variation over time in correction factors, calculated using Equation (4.2), is that of the Canberra measure.

4.4.4 Constrained Optimization Over Time

For the distance measure associated with the smallest variance, a new set of time constant correction factors $(\mathbf{r}, \mathbf{s})$ are estimated. This is undertaken by generalizing the Canberra distance function (which had the smallest variation) for an array of migration tables over time,

$$f(r_j, s_i|m_{ijtk}) = \sum_{i,j,t} \frac{|r_j m_{ijt1} - s_i m_{ijt2}|}{r_j m_{ijt1} + s_i m_{ijt2}}. \tag{4.5}$$

Thus estimates are based on a number of distance measures for each origin-destination combinations over a series of annual flow tables. This optimization was undertaken with constraints on correction factors for data rated as good by Erf (2007). As in the previous section, unknown counts were used to adjust reported data, ignoring inter-Nordic flows and using the `nlminb` function. The resulting estimates of correction factors are given in Table 4.5. Comparisons of these values with past estimated correction factors estimates are difficult due to different constraint systems used. However, their values can be considered in general terms by their relation to unity. Correction factors greater than one result in an increased scaling of reported counts, whereas values lower than one result in a decreased scaling. Past estimated correction factors for the countries in Table 4.5, such as from Poulain (1993) or Poulain and Dal (2008) have similar effects. Notable exceptions are the values of Luxembourg for which Poulain and Dal (2008) estimated receiving and sending correction factors to be 0.991 and 1.194. These differences may be explained by the allocation of counts of unknown origin and destination for each data source previous to the estimation of correction factors in Table 4.5. Differences were also found for Austrian data where sending and receiving correction factors where estimated to be 1.039 and 1.694

Table 4.5: Estimated Correction Factors for the 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

The correction factors were applied to existing data to create a series of migration flow tables using (4.3). The resulting harmonized flow values, for the data presented in Figure 4.2, are shown in Figure 4.6. The black line shows the harmonized values of y_{ijt} and red and blue lines the receiving and sending data, adjusted for unknown counts. For selected origin-destination pairs, receiving values are scaled by their country specific correction factors, r_j, when available. An example of this is shown by the reduction in German destination values, regardless of sending values, where a constant difference in the harmonized and receiving values is visible. For cells in a flow table with no receiving correction factor but in rows (from origins) with a sending correction factor, an scaling of s_i was made. An example of this process is shown for harmonized data for flows from Germany which take the same pattern as German sending data only in destinations where no receiving correction factors are present (such as to Great Britain). All other reported data is ignored, hence no black line is shown in cells such as Belgium-Great Britain as the only data available are considered unreliable.

For 2006, the estimated migration flow table of harmonized data are shown in Table 4.6. In contrast to the original reported double entry table for the same year in Table 4.1, only one value is estimated for each cell. For flows to and from countries that had correction factors constrained to unity, values are the same or have small differences from the allocation of unknown flows.

4.5 Summary and Conclusion

In this chapter a methodology for the harmonization of data for international migration flows tables was outlined. It commenced by considering a set of tables over time, with adjustments to reported data for flow values that may be reported differently through data exchanges and data dissemination problems. Comparisons of constrained optimization methods for international migration flow data was then analyzed across time. This was undertaken for data considered by experts to be of a reasonable quality, and resulted in

Figure 4.6: Harmonized and Reported Migration Counts (000's) for each available Origin-Destination Combination of EU15, 2002-2006.

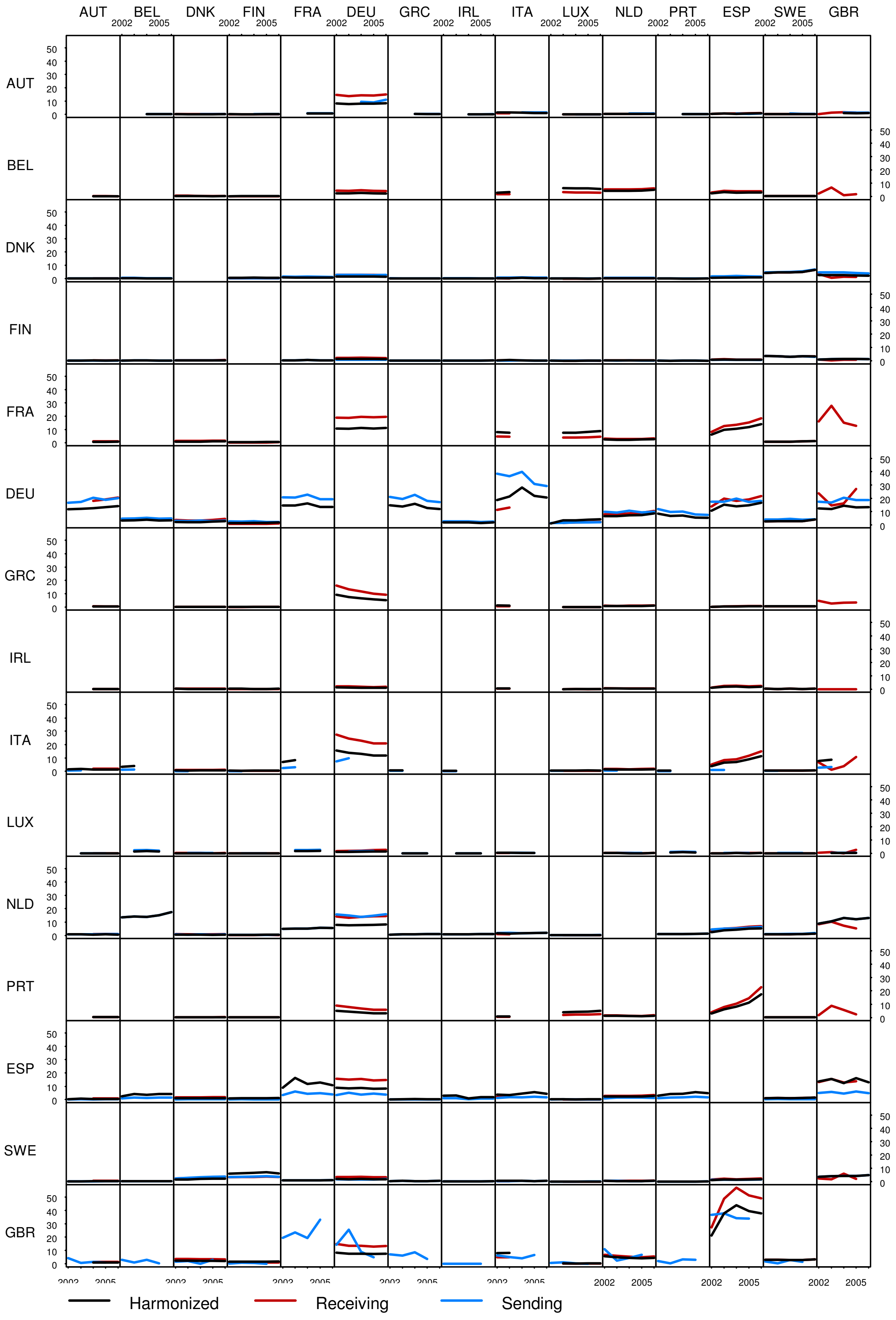


Table 4.6: Estimated Harmonized Migration Flows from each Origin-Destination Combination of the EU15 in 2006

	AUT	BEL	DNK	FIN	FRA	DEU	GRC	IRL	ITA	LUX	NLD	PRT	ESP	SWE	GBR
AUT		237	221	197	662	8540	371	143	1102	87	486	235	871	326	1002
BEL	208		340	373		2387				5870	5059		3065	432	
DNK	146	254		644	704	1487	109	158	415	165	401	79	973	6470	2174
FIN	234	187	293		335	1151	47	241	199	30	276	24	696	3110	1193
FRA	803		1128	744		11079				8993	2762		14104	1223	
DEU	14279	3485	2875	1790	13654		12014	1788	20575	4415	8576	5384	16647	4114	13293
GRC	285		162	147		5197				39	967		627	644	
IRL	114		174	241		1000				136	387		1782	312	
ITA	1301		671	528		11679				563	1617		11411	794	
LUX	60		99	96		1515					140		143	84	
NLD	666	17475	604	496	5591	8154	944	1038	2069	282		1429	5378	1327	13144
PRT	247		162	123		3272				5015	1395		17448	212	
ESP	586	4111	1186	1281	10834	8250	331	1946	4485	242	2774	4814		1468	13103
SWE	470	452	2333	6290	1267	1846	724	355	612	106	647	158	1900		5041
GBR	994		2080	1824		7486				339	4568		37850	3314	

fairly stable correction factor estimates between each time period. The Canberra distance measure was identified as having the lowest variance across time periods, and thus was generalized for the calculation of a single correction factor for each data set over a series of migration flow tables. These factors were used to scale a series of migration data that experts believed to be of reasonable quality at a harmonized level.

The constrained optimization techniques used in this chapter estimated an incomplete set of data that can be used within an overall methodology for estimating international migration flow tables of comparable data. In reference to the desirable criteria in Table 1.1, estimates can be considered consistent and reliable. Data sources that were ranked with scores of good for timing, completeness and accuracy by Erf (2007) had their correction factors constrained to unity. The remaining correction factors were applied to reported flows to allow scalings in their values towards the levels in the constrained data sources. Reliability in some data sources was ensured by preserving the original receiving data distributions of reasonable data and constraining good data. Other desirable criteria suggested in Table 1.1 for the estimation of international migration flow tables of comparable data, namely, completeness and an associated precision measure, were not obtained in this chapter. However, these criteria will be addressed in the next two chapters of this thesis.

The methodology presented allows expert opinion to determine which data sources can be considered of reasonable quality and which should be ignored and treated as missing. Expert opinion may also help in the treatment of counts that have an unknown origin or destination, as will be discussed later in this section. The methodology can be relatively easily replicated. Alternative routines such as `fmincon` in Matlab produced very similar results to those from S-Plus presented in Table 4.4. The portability of the methodology to different countries is dependent on the availability of reliable data and expert opinions that are comparable over multiple data sources.

The harmonization methodology in this chapter relies on a number of assumptions. First, prior to the application of a constrained optimization procedure, counts of migrants with unknown origins or destinations were distributed evenly across all countries. This procedure assumes that the location of the future or past residence is independent of the missing process. This could potentially be untrue in some countries, where counts from or to origins or destinations might be more likely to be unavailable than others. Further expert opinion on data sources could help avoid such an issue. Second, constraints were placed on sending data with the assumption that a ranking score of good for all data characteristics is the same as a ranking of good for receiving data. However, it is generally considered that receiving data is of a better quality than sending data due to the difficulties in tracking migrants leaving a destination in comparison to arrivals. Third, it was assumed that measurement systems in all countries with estimated correction factors remained unchanged throughout time. In the case of Spain this might have been over optimistic, as a volatile pattern in the counts of unknown origins or destinations is present. However, most counts to specific origin and destinations are fairly stable over the time period and the literature considered (Breem and Thierry, 2006b) suggests that changes in the measurement and collection occurred previous to the studied time period.

Future research on the methods used in this chapter may further enhance estimates and the methodology. Improved correction factors could be obtained using data from a longer period, such as reported counts previous to 2002. Efforts were made to attain a longer series of origin-destination flow data collected by Eurostat in the Joint Statistical Questionnaire on Migration. Problems appeared in the validity of reported sending values which, when entered in a migration flow table, produced vertical patterns. After correspondence with the Eurostat support office, it was discovered that this unusual pattern was caused by the deletion of some data. If available, a greater amount of information could be incorporated into the estimation of correction factors, given the assumption that migration data collection methods by national statistics institutes remained unchanged. More data could also be helpful in detecting flows that have large amounts of variation in comparison with other collection sources. Plots of comparative flows across time, as seen in Figure 4.1, allowed the easy discovery of some questionable data sources such as those provided by Great Britain. With a longer series of data these plots could help inform users as to which data sources are eligible for the estimation of a correction factor to scale their data. In this chapter, recognition of reliable sources was taken from a single report of Erf (2007) which created a quantitative representations of data collection techniques. Further work in this area could have the potential to incorporate such measures into distance functions. For example, a weighting of distance measures could be implemented to reflect different timing criteria used in each data source.

The use of non-linear optimization routines in statistical software allowed a great deal of flexibility to change constraints and distance measures. Plots of correction factors provided a number of useful indicators to the performance of optimization routines, constraint sets and the effect of ignoring specific flow values. Further alterations to these manipulations could be studied such as introducing more realistic bounds for correction factors from expert opinion. Final estimated correction factors in Table 4.5 differed from previous estimates from alternative methodologies, although comparisons are difficult to make due to different constraints and data used. Final receiving correction factors tended to be lower than those of Poulain and Dal (2008). This is partly driven by the exclusion of inter-Nordic flows and the lack of a constraint on total flows. If required, correction factors could be altered either directly through constraints or indirectly though estimation boundaries. For example, if an expert judges the level of under-counting of receiving data in Finland, a new constraint, different from one, could be imposed to reflect the missing percentage. Alternatively, tighter bounds in the parameter space could force estimates to be in the neighbourhood of those supplied by expert opinion. Routines might also be easily constrained to harmonize data to an alternative set of countries that may use different timing criteria in their migration definition, such as a six month definition, as used by multiple migration data sources in the EU15.

In conclusion, differences in available migration data from reliable reporting countries represent a measure of inconsistencies between reporting data sources. By scaling counts using correction factors, these distances can be minimized resulting in a harmonized data set. There exist multiple methods to measure these distance and strategies to minimize their overall levels. In this chapter, a non-linear optimization routine was used which allows boundaries to be easily set to constrain the parameter estimates. Using sub-tables composed on reliable data, as informed by expert opinion on various aspects of the data collection process, correction factors were estimated in multiple time periods. This proved a useful exercise, allowing the best distance measure to be determined for the estimation of time constant correction factors across a series of migration tables. After applying these correction factors the resulting data have the potential to be studied in relation to covariates factors suggested by international migration literature. As demonstrated in the next chapter, model based methods can be used to allow imputations for missing data.

Chapter 5

Estimating Missing Data in International Migration Flow Tables

5.1 Introduction

In this chapter, model based imputations for missing data in flow tables are derived. International migration flow table often contain missing data, creating difficulties in the analysis of population behaviour and change. Data may be missing for a number of reasons. First, national statistics institutes in some countries do not provide reported counts due to the lack of a data collection infrastructure. Second, international migration flow data tends to be collected to meet a domestic demand. Flows to or from certain countries, that are not of interest to their governments, might not be measured. Third, some countries may have insufficient data collection methods to report migration by origin or destination. For example, in Great Britain the International Passenger Survey (IPS) is used to help provide international migration flow data. Carried out during border crossings to and from Great Britain, estimates for the origin or destination or migrants where low volumes of movements exist are inadequate (Perrin and Poulain, 2006b). Finally, 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 statistics institutes not providing, or the organizations not publishing data, despite collection procedures being in place.

As migration flows can potentially be counted by both sending and receiving countries incompleteness for some cells in a double entry migration table may not always be problematic. When data is not collected by one of these sources, the partner country may provide an adequate estimate for the flow value. If the reporting partner's data is believed to be of good quality but uses alternative methods or definitions, there exists the possibility that estimates can be scaled to a given definitional requirement, as discussed in the previous chapter.

Alternative methods to impute missing cell values into international migration flow tables have been ad-hoc (see for example Raymer (2007) or Poulain and Dal (2007)). Although not without value, there exists a limited amount of research into their theoretical properties. A more comprehensive understanding of imputations techniques can be explored using statistical methods based on likelihood theory for analysis with missing data. One such method is the Expectation-Maximization (EM) algorithm of Dempster et al. (1977), a general purpose routine for maximum likelihood estimation.

In order to maximize the likelihood, a distributional assumption regarding the data is required. This typically takes the form of a statistical model which describes the behaviour of a random variable, such as a migration flow. Models for migration flow tables reside predominantly in internal migration research, for which a range of distributional assumptions have been explored (see, for example, Congdon (1991). This chapter intends to use similar models for the modelling of incomplete international migration flow tables. Using covariate information drawn from international migration theory, imputations for missing data are derived using the EM algorithm.

This chapter commences by reviewing models for population mobility tables, which have been found to have statistical equivalences to generalized linear models (introduced in Chapter 2). The following section outlines the EM algorithm. Models are then fitted by implementing the algorithm on the harmonized migration flows for the EU15 between 2002 and 2006 (see previous chapter) in order to account for the missing data. This new application of a popular statistical missing data technique allows imputations for missing cell values often found in international migration flow tables.

5.2 Models for Migration Flow Tables

Flowerdew (1991) outlined two main approaches to the analysis of flow tables that are commonly used for internal mobility data: the gravity model and the spatial interaction model. The gravity model approach derives from movements between regions in a similar manner to particle responses to two gravitational masses, as proposed by Newton in Principia Mathematica. Stewart (1941) and Zipf (1942) framed this approach for migration data, relying on statistical estimation of migration levels, given information on each origin, destination and a measurement of interactions between them. The spatial interaction models, associated with Wilson (1970) are based on mathematical algorithms to calibrate a constrained model to origin and destination totals. There are numerous formulations of spatial interaction models such as bi-proportional adjustment, information gain minimizing and entropy maximizing which include various constraints and interaction terms (Willekens, 1983).

Poisson regression models have become a popular method for representing migration models as they relate gravity and many spatial interaction models in a single comparative framework. Flowerdew (1982) and Willekens (1983) showed that a Poisson regression model with either row or column dummy covariates are equivalent to an origin or destination constrained spatial interaction model, and where both covariates are present, a doubly constrained spatial interaction is obtained. Such representations, with only categorical covariates, are also known as log-linear regression models of Birch (1963). When row or column dummy covariates are not included, but other origin and destination specific factors are, a gravity model with an assumed Poisson distributed response is represented (Flowerdew, 1991).

As explained in Chapter 2, Poisson regression models are part of a range of statistical models known as generalized linear models of Nelder and Wedderburn (1972), which link together a number of models that relate a random response variable to a systematic linear predictor. This statistical formulation of a migration table has several important advantages over more traditional approaches. Willekens (1983) noted that Poisson regression models enhance the structural analysis of spatial interaction, have greater clarity and simplification of parameter estimation and open the opportunity to apply a wide range of statistical theory. Guy (1987) expanded upon this final point for all Poisson regression models, noting the ability to provide standard diagnostics and better model specification. In addition, non-specialist statistical software may be used to fit generalized linear models using efficient algorithms for obtaining maximum likelihood parameter estimate. These also have greater flexibility for alternative functional forms to extend models beyond conventional size and distance variables and with a choice of error specifications.

Flowerdew and Aitkin (1982) noted some drawbacks in implementing Poisson regression models to migration flow tables. Arguably, the most prominent of these was an inability to provide an adequate fit to data. Previous attempts to fit log-linear models, such as that of Flowerdew and Lovett (1988) and Flowerdew (1991), showed that the best fitted models contained origin and destination (or table row and column) covariates. Despite adding further interaction-based explanatory factors, which improved model fits, the remaining deviances of models were still deemed unsatisfactory. The lack of fit was attributed to the equivalence of the first and second moment in a Poisson distribution. The use of a single parameter distribution assumed each movement from a given origin to a destination occurred independently, having controlled for explanatory factors. However, data in origin-destination tables are aggregated over individual characteristics. Congdon (1991) noted that without the ability to disaggregate data by more individual level factors, such as migrant age or sex, Poisson regression models fit poorly.

One solution to this problem was to fit a linear regression to the logarithm of migrant counts. Flowerdew and Aitkin (1982) noted this approach had a number of problems when fitted to migration count data. First, the introduction of the logarithmic scale creates a bias in the estimate of the mean when the antilogarithm was taken. Consequently, wrongly signed or insignificant coefficients may be included in a model. Second, a log-normal assumption for a count response has a theoretical dissatisfaction of modelling a discrete valued process by a continuous distribution. Finally, a log-normal regression model presupposes a common variance for mobility table data where there is often a wide variation in cell values. Davies and Guy (1987) suggested three alternative solutions for when a Poisson assumption in mobility tables was violated: a parametric approach using negative binomial regression model, a quasi-likelihood approach of introducing a new parameter for the mean-variance ratio and a pseudo likelihood approach of estimating a variance-covariance matrix of parameter estimates given a misspecified model. In this chapter the former of these three is further explored as its parameter estimates are based on full likelihood methods. This allows missing data techniques such as the Expectation-Maximization (EM) algorithm to be fully utilized under a negative binomial assumption for a response variable.

5.3 The Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative algorithm for maximum likelihood estimation in incomplete data problems. Used in multiple statistical settings, the EM algorithm is a prominent tool in estimation when there are missing data on random variables, such as the number of migrants between two countries, whose realizations would otherwise be observed. Developed by Dempster et al. (1977), the motivating idea behind the EM algorithm is to augment the missing parts of a data set with temporary values to complete the data and allow the estimation of model parameters to proceed in a cycle of simple estimation steps. Each cycle of the EM algorithm consists of two steps.

1. If we let $\boldsymbol{\theta}^r$ denote the current guess of the parameters at iteration r, y_o be the observed data and z denote the missing data to be augmented. The E-step (expectation step) finds the expected augmented log-likelihood $Q(\boldsymbol{\theta})$ if $\boldsymbol{\theta}^r$ were $\boldsymbol{\theta}$. This can be expressed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r) = E(l(\boldsymbol{\theta}|y_o, z)|y_o, \boldsymbol{\theta}^r) \tag{5.1}$$

   where $l(\boldsymbol{\theta}|y_o, z)$ is the log likelihood of $\boldsymbol{\theta}$ given the augmented data.

2. The M-step (maximization step) determines $\boldsymbol{\theta}^{r+1}$ by maximizing the expected augmented log-likelihood.

The algorithm is iterated until $\left|\left|\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r\right|\right|$ or $\left|\left|Q\left(\boldsymbol{\theta}^{r+1}\left|\boldsymbol{\theta}^r\right.\right) - Q\left(\boldsymbol{\theta}^r\left|\boldsymbol{\theta}^r\right.\right)\right|\right|$ is sufficiently small, and hence a maximum of the augmented log-likelihood is reached.

The EM algorithm has a number of appealing proprieties relative to other iterative algorithms for finding maximum likelihood estimates. Little and Rubin (2002, p167) noted that the EM algorithm is numerically stable as each iteration increases the likelihood, has fairly reliable convergence and often easy to program as no evaluation of the observed likelihood nor its derivatives are involved. In addition, the M-step can be easily implemented in standard statistical software by performing a fit to the current complete data at each step. Mclachlan and Krishnan (1996, p33-4) noted associated problems of the EM algorithm including: a slow convergence rate which may occur when there is a large fraction of missing data, the lack of a built-in procedure for producing an estimate of the covariance matrix of parameter estimates and the lack of a guaranteed convergence to a global maximum. The first two problems can be alleviated by choosing appropriate starting values and using a supplementary methodology that will be further explored in the following chapter. The possibility of not converging to a global maximum is a problem faced by all optimization algorithms and the EM algorithm is no different in this respect. In some cases this can be alleviated by using multiple starting points, as used throughout the remainder of this thesis, to check that the maximum reached is not localized. There exist other procedures such as simulated annealing to tackle more intricate situations which tend to be complicated to apply, as discussed by Little and Rubin (2002, p167).

5.4 Modelling Incomplete International Migration Flow Tables

In this section, negative binomial regression models are fitted to incomplete international migration flow data for the EU15 countries, presented in Figure 4.6. In order to account for the missing data, model parameters are estimated using a EM type algorithm. In keeping with statistical modelling, the harmonized data are treated as observed values.

Of the 1050 cells (made from a 15 × 15 × 5 non-diagonal mobility table over 5 time periods), 819 have observed, harmonized values. In the 210 flows for which reported counts could potentially be produced, 30 had no observations of harmonized data in any years. This was greater than the actual reported data (20), as some values are ignored due to their poor quality.

A function was written in S-Plus to obtain estimates of parameters in a negative binomial regression models using a EM type algorithm (shown in the Appendix). The function requires a fitted model object of class `negbin`, for which the model matrix of the specified model is utilized in the M-step. This can be obtained by fitting a proposed model using the `glm.nb` function of the MASS library (Venables and Ripley, 2003) and omitting any missing data. Given the model matrix, parameter estimates are generated in the routine by augmenting the missing flow counts with temporary values. These values are estimated in the M-step of the algorithm using the `glm.nb` function. The routine continues until the specified stopping criteria are met. Included in the output is a record of parameter and imputations at each iteration.

Note, this routine is not a true EM algorithm as in the M-Step it maximizes the negative binomial likelihood augmented with expected values of the missing data at each step. This is opposed to taking the expectation of the augmented likelihood as presented in the previous section. Hence, in the estimation of the parameters, the correct $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r)$ is not maximized . This is because the in the correct $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r)$ requires the expectation of the $\log(\frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})})$ (derived from the first terms in the probability density function shown in (2.34)).

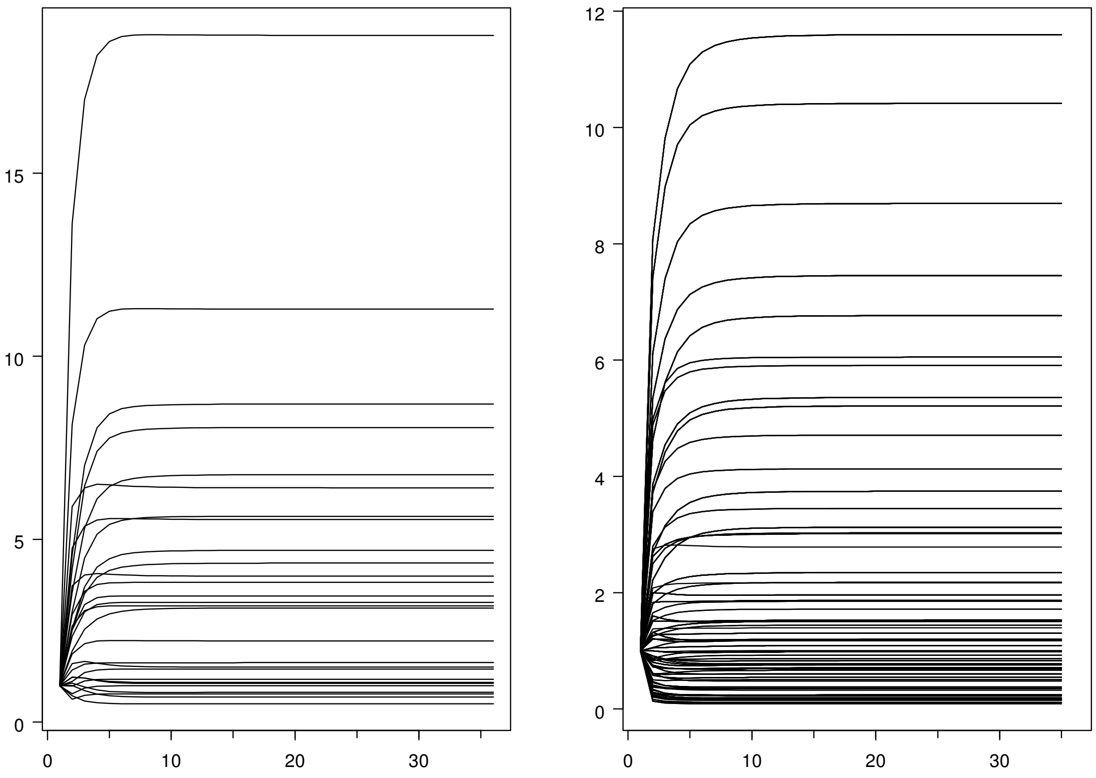
As a result of using the routine in the Appendix, derived estimates of α are not maximum likelihood estimates and may further impact the calculations of other model parameters and their associated variances. Further work is required to better understand these effects (which are believed to be minor), including using a more suitable EM type algorithms. For example, the Monte Carlo EM algorithm of Wei and Tanner (1990) avoids the direct calculation of the expected augmented likelihood in the E-Step by simulating missing values from their condition distribution to provide maximum likelihood estimates of parameters.

Initially, a spatial interaction model that was equivalent to a quasi-independent model was fitted. This can be specified in a similar manner as (2.35),

$$\log \mu_{ijt} = \beta_1 + \beta_i^O O_i + \beta_j^D D_j, \tag{5.2}$$

where $\mu_{ijt} \equiv E(y_{ijt}|\boldsymbol{\beta}, \alpha, \mathbf{x}_i^T)$ and β_1 is a constant parameter for the baseline category, β_i^O the set of 14 origin parameters and β_j^D the set of 14 destination parameters, corresponding to origin O_i and destination D_j respectively. As previously mentioned, spatial interaction models give superior over gravity models but at the cost of aliasing out additional origin and destination effects. They also allow a level of basic measure in the overall attractiveness of countries for migrants moving to and from each nation to be obtained. Such measurements are commonly referred to as push and pull factors (Lee, 1966).

Figure 5.1: Exponentiated Covariate Parameter Estimates, $\boldsymbol{\theta}$ (left) and Missing Data Values, z in 000's (right) of Quasi-Independent Fit



All parameters are successfully estimated using the EM algorithm. Final estimates, and standard errors (from the observed data) are shown in the first two columns of Table 5.1 (other columns are discussed later in this section). Figure 5.1 shows a trace of the iterative estimates from the EM algorithm of this model (right plot), alongside the imputed values

Table 5.1: Mean Parameter Estimates from EM algorithm

Parameter	Spatial Interaction		Main Effects		French Interaction	
	exp(β)	se(β)	exp(β)	se(β)	exp(β)	se(β)
Constant	212.7620	0.1688	0.0007	0.6737	0.0001	0.8489
Origin:						
BEL	4.3495	0.1852	1.2615	0.1252	1.1555	0.1418
DEU	18.7707	0.1582	7.8271	0.8397	0.0460	2.5020
DNK	1.1711	0.1583	1.5461	0.1428	1.6387	0.1642
ESP	5.5379	0.1582	3.1334	0.4991	0.3182	1.2018
FIN	0.7756	0.1583	0.4540	0.1725	0.5433	0.1897
FRA	6.7664	0.1852	4.6928	0.7352	9.6826	3.7769
GBR	8.6948	0.1852	14.5710	0.7570	0.3648	1.7986
GRC	1.0894	0.1853	1.0140	0.1139	0.8751	0.1423
IRL	0.6916	0.1854	0.8422	0.2014	1.0365	0.2349
ITA	3.4470	0.1701	2.5763	0.6758	0.0886	1.7146
LUX	0.5038	0.1667	0.5314	0.8230	1.1143	0.8277
NLD	3.9895	0.1582	3.1579	0.2867	1.5833	0.3779
PRT	3.1251	0.1852	0.2257	0.2706	0.2041	0.2722
SWE	3.1787	0.1582	1.8473	0.1521	1.4836	0.1506
Destination:						
BEL	4.6995	0.1867	0.7024	0.1193	0.6172	0.1362
DEU	11.2890	0.1582	0.1182	0.8379	0.0008	2.5082
DNK	0.9953	0.1583	2.4038	0.1439	2.5266	0.1651
ESP	6.4067	0.1582	0.8533	0.4993	0.0907	1.2055
FIN	1.4570	0.1583	6.7173	0.1740	8.2222	0.1919
FRA	5.6295	0.1866	0.1655	0.7329	0.0017	1.8192
GBR	8.0547	0.1866	0.5504	0.7529	0.0151	1.8029
GRC	1.0608	0.1868	2.0992	0.1069	1.7619	0.1379
IRL	0.8179	0.1869	4.2119	0.2068	5.1961	0.2404
ITA	3.2706	0.1710	0.4308	0.6750	0.0161	1.7195
LUX	1.5046	0.1619	1.8496	0.8655	3.2087	0.8686
NLD	2.2212	0.1583	0.5530	0.2825	0.2988	0.3763
PRT	1.6291	0.1867	5.4726	0.2283	5.0024	0.2319
SWE	3.8221	0.1582	1.7893	0.1485	1.4742	0.1471
Main Effects:						
GNI			6.7608	0.3076	6.8078	0.2997
GDP			2.3310	0.3688	2.2718	0.3540
Trade			1.3431	0.0306	1.3704	0.0322
Euro			1.4333	0.0970	1.3838	0.0934
Stock			1.8190	0.0192	1.8641	0.0196
French			3.3312	0.1718	1.9649	0.1884
English			0.4017	0.2957	0.5557	0.2957
Population					0.9647	0.0142
Time					1.0689	0.0319
French Origin Interaction:						
GNI					3.8759	0.4917
Euro					0.5698	0.1750
Population					1.0183	0.0052
Stock					0.7305	0.1079
Distance					0.5696	0.3278
French Destination Interaction:						
Stock					125.6465	0.8586

for 231 missing cell values (left plot). An initial value of one was chosen for all parameter estimates whose values all met a convergence criteria of $\left|\left|Q\left(\boldsymbol{\theta}^{r+1}\left|\boldsymbol{\theta}^{r}\right.\right)-Q\left(\boldsymbol{\theta}^{r}\left|\boldsymbol{\theta}^{r}\right.\right)\right|\right| < 10^{-5}$ after 36 iterations.

The exponentiated origin parameter estimates from the quasi-independent model measure the level of attraction over the entire time period, in comparison to Austria, which was used as a reference category. Values varied from 18.7707 and 8.6948 for Germany and Great Britain to 0.5038 and 0.6916 for Luxembourg and Ireland, respectively (with reference to unity for Austria). Exponentiated destination parameter values (where Austria was again the reference category) varied from 11.2890 and 8.0547 for Germany and Great Britain to 0.8179 and 0.9953 for Ireland and Denmark, respectively. The dispersion parameter was estimated as 1.2863 (using the `glm.nb` function), which is equivalent to the inverse, 0.7774 for α in equation (2.34). A Z-test provided strong evidence that $\alpha > 0$, suggesting the data was overdispersed and hence the negative binomial model was more appropriate than an equivalent Poisson model.

5.4.1 Additional Information

In order to provide more reasonable imputations, the quasi-independent model was expanded upon. There are many theories that explain international migration, see for example Massey et al. (1993) or Greenwood and Hunt (2003). Data for economic, geographical and demographic factors suggested by these theories are often comparable across multiple nations and available from data bases of international organizations. Data on nine of these factors were chosen. Where possible, information across time was taken to help reflect trends in migration flow counts seen in Figure 4.6.

Four covariates on economic systems were constructed: the origin-destination ratio of Gross National Income (GNI) per capita 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`). Measures with a purchasing power parity adjustment, to account for differences in relative living costs and inflation, were used. A per capita measure for GNI was taken to reflect a macro measurement of the differences in wages between origins and destinations. GDP was measured on a national level (rather than per capita) to reflect differences in economies income and output. The logarithm of this ratio was taken due to the high 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 may 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 and 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 and 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. A covariate for population was used to control for higher migration flows between countries with large populations such as Germany and France. For each flow, a measure from the sum of origin and destination populations was calculated. Hence, the same covariate value is obtained regardless of the flow direction. Data was 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 data base from the 2000 round of population censuses. Covariates on languages were considered to further reflect social and linguistic similarities. These were 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 (from 0.87 × 1.00) and 0.0667 (from 0.29 × 0.23) 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 and correlation amongst repeated counts of the same origin-destination pair, over the time period.

5.4.2 Main Effects Model

In order to attain a better model fit and more realistic imputation the Akaike Information Criterion (AIC) was used to select the most suitable variables for a main effects model.

$$AIC = -2l(\boldsymbol{\theta}|y_o) + 2p, \tag{5.3}$$

where $l(\boldsymbol{\theta}|y_o)$ is the log likelihood of $\boldsymbol{\theta}$ given the observed data, y_o, and p is the dimension of $\boldsymbol{\theta}$. Comparisons of potential models were undertaken using the `stepAIC` function in the MASS library (Venables and Ripley, 2003). The function operates by examining the inclusion of potential covariates by their contribution to the AIC of the model, performing a stepwise search in both directions, adding and dropping variables. Included in a pre-condition in the scope of models to be searched were origin and destination covariates. The final model selected by the `stepAIC` function can be specified as,

$$\begin{aligned}\log \mu_{ijt} &= \beta_1 + \beta_i^O O_i + \beta_j^D D_j \\ &+\beta_2 GNI_{ijt} + \beta_3 \log GDP_{ijt} + \beta_4 \log TRADE_{ijt} + \beta_5 EURO_{ij} \\ &+\beta_6 \log STOCK_{ij} + \beta_7 FRENCH_{ij} + \beta_8 ENGLISH_{ij}\end{aligned} \tag{5.4}$$

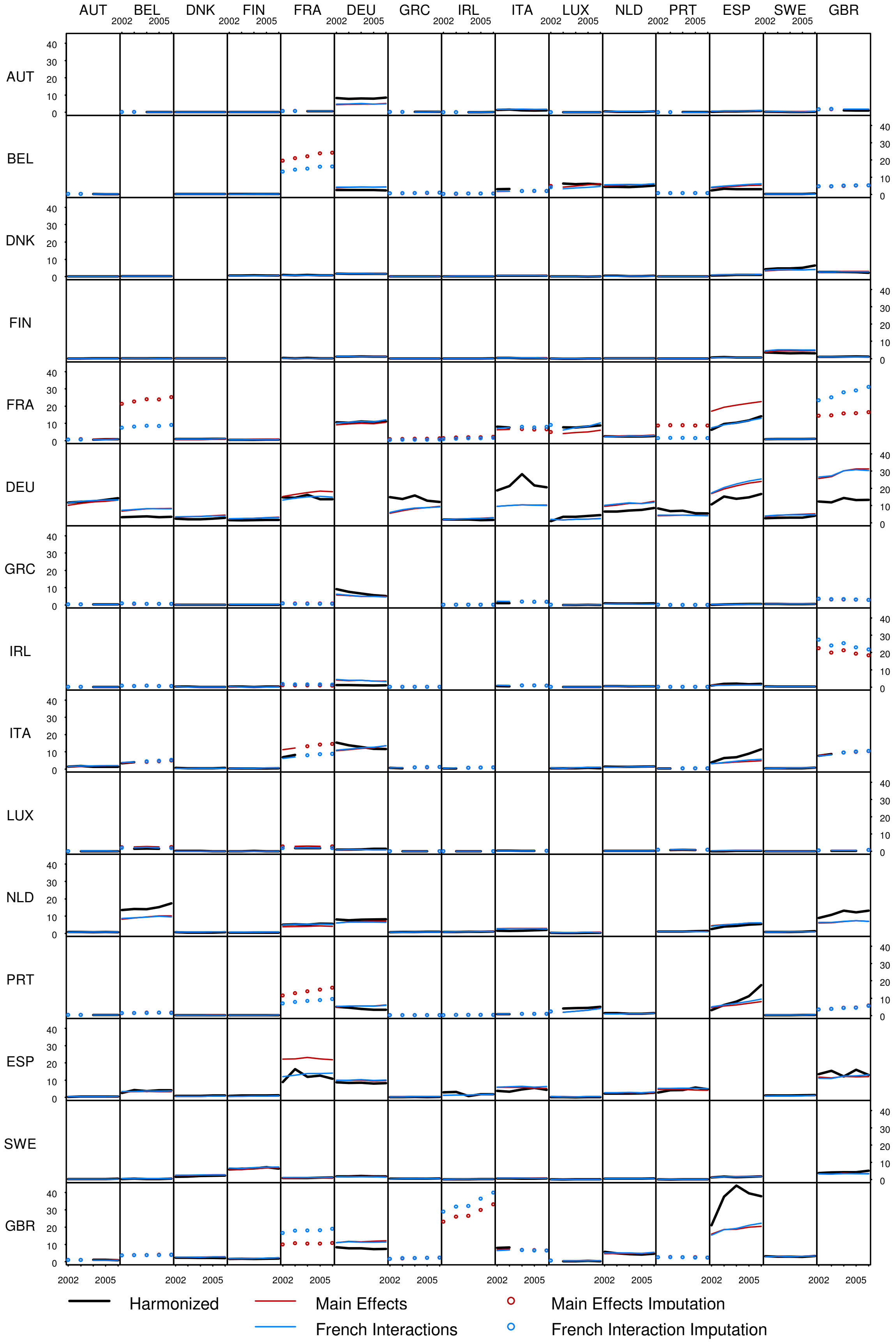
where all covariates are flow specific, and GNI, GDP and trade were time-varying. Covariates for distance, contiguity and the German language were found to be ineffective in reducing the AIC. The remaining covariates were included in the final main effects model which had an AIC statistic of 12,101 in comparison to 13,548 of the quasi independent model (shown in Table 5.2). Convergence when fitted with the EM algorithm was obtained after 43 iterations with stopping criteria of 10^{-5}. Fitted values are shown by the solid red line where original data existed, and by red marks for the imputations on previously missing data in Figure 5.2.

Table 5.2: Dispersion Parameter Estimates from EM algorithm and AIC

	Spatial Interaction	Main Effects	French Interaction
α	0.7774	0.1557	0.1432
se(α)	0.0573	0.3168	0.3460
AIC	13547.69	12101.26	12047.97

Parameter estimates for the selected covariates are shown in the third and forth columns of Table 5.1. Origin and destination effects strayed from their values found in the quasi-independent as additional factors were controlled for. The estimated exponentiated parameters effects for economic factors (6.7608 for the ratio of GNI per capita, 2.3210 for the logarithm of the ratio of GDP, 1.3431 for the logarithm of trade volume and 1.4333 for Euro region), logarithm of migrant stocks (1.8190) and French prevalence (3.3312) 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 (0.4017) 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 is 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 throughout most origin-destination pairs. The dispersion parameter was estimated to be 0.1557 with standard error 0.3168 (as displayed

Figure 5.2: Data and Main Effects Model Fits of Migration Flows (000's) from each Origin-Destination Combination of EU15, 2002-2006.



in Table 5.2). This was noticeably smaller than the quasi-independent fit indicating evidence for a control on overdispersion in the main effects models.

5.4.3 Interaction Models

To gain a further superior fit the `stepAIC` function was run once more with an extended scope of models to consider all two-way interactions, with one exemption, 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 fitting 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. This resulted in producing a total of 243 new parameters (not shown in Table 5.1). Many of these parameters were unidentified and imputations were unreasonable as large shares of available information in the observed data are used to fit model parameters for the complete data. Consequently, the observed value all had extremely good fits in all years (not shown on Figure 5.2). This was reflected by AIC statistics as low as 11,072 (not shown in Table 5.2) a large reduction from the main effects model (12,101). In addition to these problems, different parameter values are estimated for different starting values. This is also due to the lack of observed data relative to the number of parameters

Whilst a single model with many interactions and multiple parameters may not be plausible for a migration table involving many countries, interactions for single countries can be constructed to improve model imputations where deemed necessary. Analysis of the fits from the main effects model in Figure 5.2 showed reasonable imputations for most previously missing cells. Clear exceptions are 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 to model, where effects of some factors may vary substantial for migration flows to or from individual nations.

A closer fit was obtained by considering interactions for the 11 covariate parameters (including 3 languages effects) outlined in Subsection 5.4.1 with France as both an origin and destination. The 22 additional covariates where considered by the stepwise model fitting algorithm. The final model selected by the `stepAIC` function was,

$$\begin{aligned}
\log \mu_{ijt} \quad = \quad & \beta_1 + \beta_i^O O_i + \beta_j^D D_j \qquad (5.5)\\
& + \beta_2 GNI_{ijt} + \beta_3 \log GDP_{ijt} + \beta_4 \log TRADE_{ijt} + \beta_5 EURO_{ij} \\
& + \beta_6 \log STOCK_{ij} + \beta_7 FRENCH_{ij} + \beta_8 ENGLISH_{ij} + \beta_9 POP_{ijt} \\
& + \beta_{10} TIME_t \\
& + \beta_{11} O_{FRA} : GNI_{ijt} + \beta_{12} O_{FRA} : EURO_{ij} + \beta_{13} O_{FRA} : POP_{ij} \\
& + \beta_{14} \log O_{FRA} : STOCK_{ij} + \beta_{15} O_{FRA} : DIST_{ij} + \beta_{16} \log D_{FRA} : STOCK_{ij},
\end{aligned}$$

where covariates are flow specific, and are equal to zero in non-French rows (or columns) corresponding to interaction terms with France as a origin (or destination). The AIC of final selected interaction model was 12,047 (as displayed in Table 5.2) a further reduction in comparison to the main effects model but with more parameters (from 37 to 45), see Table 5.1. Of these, six were new interaction covariates and two more main effects for population and time. Additional main effect covariates are included, partly as higher level interactions with other covariate with France were effective and hence its main effects are also useful. Alternatively, these have been included as parameter estimates from the original main effects model are altered by the inclusion of interactions and thus more or less factors might be added to cover the change in model fit.

Of the six new interactions, five (GNI ratio, population sum, the Euro zone, stock and distance) were with France as an origin and one (stock) were with France as a destination. Their inclusion indicated evidence for different effects for a French origin (or destination) on the expected migration flows leaving (or arriving) in comparisons with a general effect for all nations. All parameters were identifiable and led to a noticeable change in the fit on flow values in the French row and column. These are shown in Figure 5.2, where fitted values are shown by the solid blue line where original data existed, and by blue marks for the imputations on previously missing data. For flows from Italy to France, imputations in later years follow neatly from harmonized data in the first two time periods. In addition, flows from Belgium, which were considered unusually high have fallen, whilst flows to and from larger countries such as Great Britain have increased. The dispersion parameter was estimated to be 0.1432 (shown in Table 5.2), again noticeably smaller than the previous model, indicating further control on overdispersion in the interaction model.

For 2006, the complete migration flows table is shown, where bolded values are from scaled reported flows in Table 4.6 and non-bolded values are imputations from the model. For all cells in the table there exists an estimate. The non-bold estimates are fixed in the EM algorithm, as observed values, are unchanged regardless of the model used. Estimates for cells that were previously missing are dependent on the model used to base imputations on.

5.5 Summary and Discussion

In this chapter, a complete set of estimates of international migration flow tables are created, using a spatial interaction model fitted using the EM algorithm on the harmonized flows from the previous chapter. The choice of model was, for the most part, left to an stepwise model selection program, although some form of expert opinion was used to allow consideration of further parameters for models where imputations were initially unreasonable.

In reference to the desirable criteria in Table 1.1, estimates for the international migration tables in Figure 5.2 are now considered complete, consistent and reliable. Completeness and consistency of estimates was achieved using the EM algorithm to fit negative

Table 5.3: Estimated Migration Flows from each Origin-Destination Combination of the EU15 in 2006 using the French Interaction Model

Bold values are from scaled reported flows in Table 4.6, non-bold values are imputations.

	AUT	BEL	DNK	FIN	FRA	DEU	GRC	IRL	ITA	LUX	NLD	PRT	ESP	SWE	GBR
AUT		**237**	**221**	**197**	**662**	**8540**	**371**	**143**	**1102**	**87**	**486**	**235**	**871**	**326**	**1002**
BEL	**208**		**340**	**373**	16135	**2387**	945	555	1965	**5870**	**5059**	773	**3065**	**432**	5189
DNK	**146**	**254**		**644**	**704**	**1487**	**109**	**158**	**415**	**165**	**401**	**79**	**973**	**6470**	**2174**
FIN	**234**	**187**	**293**		**335**	**1151**	**47**	**241**	**199**	**30**	**276**	**24**	**696**	**3110**	**1193**
FRA	**803**	9166	**1128**	**744**		**11079**	674	1623	8022	**8993**	**2762**	1498	**14104**	**1223**	31041
DEU	**14279**	**3485**	**2875**	**1790**	**13654**		**12014**	**1788**	**20575**	**4415**	**8576**	**5384**	**16647**	**4114**	**13293**
GRC	**285**	595	**162**	**147**	580	**5197**		108	1747	**39**	**967**	42	**627**	**644**	2860
IRL	**114**	579	**174**	**241**	1314	**1000**	118		729	**136**	**387**	145	**1782**	**312**	21620
ITA	**1301**	5220	**671**	**528**	8811	**11679**	1212	1002		**563**	**1617**	462	**11411**	**794**	10527
LUX	**60**	1803	**99**	**96**	1780	**1515**	70	50	185		**140**	786	**143**	**84**	577
NLD	**666**	**17475**	**604**	**496**	**5591**	**8154**	**944**	**1038**	**2069**	**282**		**1429**	**5378**	**1327**	**13144**
PRT	**247**	1433	**162**	**123**	9407	**3272**	80	292	970	**5015**	**1395**		**17448**	**212**	5436
ESP	**586**	**4111**	**1186**	**1281**	**10834**	**8250**	**331**	**1946**	**4485**	**242**	**2774**	**4814**		**1468**	**13103**
SWE	**470**	**452**	**2333**	**6290**	**1267**	**1846**	**724**	**355**	**612**	**106**	**647**	**158**	**1900**		**5041**
GBR	**994**	3870	**2080**	**1824**	18938	**7486**	2350	39925	6436	**339**	**4568**	2313	**37850**	**3314**	

binomial regression models based on the harmonized flows from the previous chapter. Hence, the resulting imputations share the characteristics of the data sources that were ranked with scores of good for timing, completeness and accuracy by Erf (2007). Further checks for reliability of the estimates were taken in this chapter by comparing imputations over time. For flows to and from countries with only partially available harmonized data (Austria, Italy and Luxembourg) checks across time helped further inform the model fitting process. Using a model based imputation method based, in statistical theory, measures of estimates precision can potentially be found. One such technique to obtain these measures is the Supplemented EM algorithm of Meng and Rubin (1991) which will be further explored in the next chapter.

The techniques shown in this chapter have also addressed the suggested criteria for the methodology to be used in international migration flow table estimation. As Willekens (1994) suggested, a model based method has been used for the estimation of missing data. This allowed some substantive understanding of flows and imputations to be based on likelihood methods. For the EU15 region studied in this chapter parameter effects for the selected covariates predominantly had the expected direction suggested by international migration theory. This could be considered to further enhance the justification for estimates to be deemed reliable, as unlike more ad-hoc impaction techniques, their values are based on the relationship with factors that are believed to influence migration. Expert opinion can be used in the model building process. In this chapter model selection was in the most part left to an automated procedure, where the covariates considered were based on past literature of international migration. Alternative covariates and modelling strategies may be pursued as and when an expert deems necessary, as discussed later on in this section.

The methodology can be relatively easily replicated given replicated in S-Plus/R given the data and function supplied in the Appendix of this thesis. The use of the EM type algorithm, for incomplete migration flow tables can be applied to models for alternative international migration tables. This can include both smaller or larger tables and additional data for previous or subsequent time periods. In this study a restriction to 15 nations was used to enable effective models for flows between politically similar countries with only a few main effects. In a more diverse set of countries, political differences between nations that may influence migration, would require additional care to obtain more reasonable estimates. The EM algorithm may also be used to estimate missing cells in international migration tables based on other types of data and populations. For example, Abel (2008) modelled incomplete tables of the stock of student migrants present in nine countries spread across the globe.

When missing data was present, the success of imputations from the spatial interaction model fitted using the EM algorithm is dependent on amount of data available. As a prerequisite for a quasi-independent model, some data on the number of flows to and from each country must be present to identify all parameters. In this study, this was achieved by combining harmonized sending and receiving data from the last chapter and analyzing trends over multiple time periods. The spatial interaction model was chosen in order to provide the best fit to the data.

Better fits for spatial interaction model could be further achieved by considering alternative covariates or redefining existing ones. For example, the time covariate was considered to be continuous for ease of interpretation, but it could have been considered as a categorical factor. This would allow time-specific effects to be estimated in the same manner as origin-specific and destination-specific resulting in a superior fitting model, but at the cost of more parameters. Interactions terms between these covariates would lead to a saturation of the model but effects may not always be identifiable in incomplete data situations if no counts exist in a given time period for a given origin or destination. It is useful to note that if interest lay in controlling for specific origin-destination combinations, such as the migration flow from Great Britian to Spain, a covariate could be built to include this term and induce a better model fit in that cell. Further covariates may improve model fits. For example, large flows such as from Germany to Italy or from Great Britain to Spain were underestimated by the main effects model. These flows may have involved moves for retirements. Covariates on related factors such as climate or migrant age could be beneficial for model fits, including imputations for missing data. The negative binomial regression model proved an effective tool to deal with overdispersion of the data. The use of an alternative error assumptions, such as a Poisson distribution, would have lead to worse fitting models and non robust standard errors in the presence of overdispersion (Davies and Guy, 1987).

The inclusion of interaction covariates with multiple levels can lead to unidentifiable parameters and unrealistic imputations for unobserved cells when implementing the EM algorithm. An alternative strategy was explored by adding interaction terms only with country specific levels, where better model fits were obviously needed. This was done for flows to and from France, resulting in improved imputed values. Better fits could have also been obtained using a similar framework for other countries where expert opinion may deem imputations unrealistic. Alternatively, a more automated approach would be to consider all levels of interactions individually for inclusion into a model via a stepwise modelling approach. However, this would require a considerable amount of computations, as the number of potential models would become very large. In addition, the inclusion of further levels of interactions for countries where existing model fits are poor may not be of great use when estimating missing values. For example, country-level interaction parameters were tentatively estimated for all flows originating from Germany and all flows into Spain. The resulting model from a stepwise model selection enhanced the fits for flows from Germany and into Spain, but only slightly altered imputations in non-German and non-Spanish flows.

The use of expert opinion in selecting covariates in the modelling process may be beneficial when the harmonized data are heavily reliant on the selected distance measure. Although the distance measures studied in Chapter 4 produced correction factors that were alike in most time periods, for different data the choice of distance measure might be very influential on the estimated harmonized data. In such a case, an alternative model is likely to be selected for the imputation of missing flows by a stepwise routine. However, expert opinion can help inform the selection process if the resulting estimates for missing cells are judged to be unrealistic. As discussed, this might involve the addition of new covariates or interaction terms to help improve model fits.

Alongside modifying interactions to country-specific levels further improvements to modelling international migration flow tables could be explored. The building of models in this chapter relied upon comparisons of competing AIC calculated using the log-likelihood of the observed, rather than, complete data. As Cavanaugh and Shumway (1998) noted it is more desirable to fit a model based on the complete data for which models are originally postulated. Criteria, such as the AICcd of Cavanaugh and Shumway (1998) and KICcd 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 by implementing the Supplemented-EM algorithm of Meng and Rubin (1991) which requires further computations during the EM algorithm. This will be further explored in the succeeding chapter.

An alternative approach for modelling data across time when estimating missing data is to consider origin-destination combinations in a marginal model, which are typically fitted using the Generalized Estimating Equation of Zeger et al. (1988). Marginal models would enable the exclusion of origin and destination specific parameters, allowing more complex categorical covariates to be fitted. These methods have been used in previous panel data studies of international migration data by Pedersen et al. (2004) and Mayda (2007). Both studies used unbalanced receiving migration flow counts from the SOPEMI reports of the OECD, where procedures to handle inconsistencies in data sources are not used. The use of marginal models for imputing missing data would require more complex parameter estimation techniques in the M-step of the EM algorithm and an assumption for the correlation structure of data. Cohen et al. (2008) used a log-normal regression model (involving origin, destination, geographic and demographic factors) to project future migration between two countries. This was based on inconsistent data from 11 countries between 1960 and 2004. Missing data was not accounted for in the estimation of parameters; however, covariates to account different data sources were included to account for differences in collection and measurers. Such covariates could be identified, despite the including of origin and destination covariates, due to the unbalanced nature of the data.

Despite the common occurrence of missing data in international population mobility tables, the application of the EM algorithm is sparse. Willekens (1999) suggested the EM algorithm as a possible method to fit spatial interaction models to constrained margins. This model was further expanded by Raymer et al. (2007) where we found the EM algorithm to be equivalent to a conditional maximization given the marginal constraints. Imputations for missing cells 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) and Raymer (2007) applied more ad-hoc methods outlined in Chapter 3.

In conclusion, the EM algorithm allows missing values in international migration flow tables to be estimated. These are based on statistical assumptions and covariate information from international migration theory. There exist a number of options for building models. In this chapter, negative binomial regression models were compared using their AIC statistics. This proved an effective strategy to deal with overdispersion and help in the model selection procedure.

Chapter 6

Estimating Measures of Precision for Missing Data in International Migration Flow Tables

6.1 Introduction

In this chapter, estimates for the measures of precision of missing cells in international migration flow tables are derived. These measures allow data users to obtain a better understanding of the possible variation of missing data estimates. As demonstrated in the previous chapter, likelihood based methods for imputing missing data allow the most likely estimates to be obtained given the data. In incomplete data situations, the Expectation Maximization (EM) algorithm allows maximum likelihood estimates to be calculated. However, unlike fitting methods readily used in complete data situations (such as IRLS or the Newton optimizer), the asymptotic variance-covariance matrix for parameter estimates is not an automatic by-product of its procedure. These matrices are useful when conducting statistical inference, allowing test statistics and standard errors to be derived.

More detailed routines exist that may account for missing data in the estimation of the variance-covariance matrix. One such method is the Supplemented EM (SEM) algorithm of Meng and Rubin (1991). The SEM algorithm, unlike alternative estimation methods, such as that of Louis (1982), do not require extra analytical calculations beyond those needed to calculate maximum likelihood estimates. This property is of considerable benefit when models include many parameters, as used in spatial interaction models with multiple regions. A further contribution of the SEM algorithm is the ease to calculate the AIC complete data (AICcd) criteria of Cavanaugh and Shumway (1998), which can enable the selection of models having accounted for missing data.

This chapter commences by reviewing the convergence properties of the EM algorithm. The SEM algorithm is based upon the rate of convergence of an EM algorithm being governed by the fraction of missing information. As explained, this principle is used to find the increased variation due to missing information which is then incorporated into the estimate of the complete data variance-covariance matrix. The following section outlines the algorithm for which a measure of the missing information can be computed. A review is then given of the AICcd which uses output from the SEM algorithm in its calculation. The SEM algorithm and AICcd are then applied to the fitting and selection of main effects models using the parameters discussed in the previous chapter. This new application of statistical techniques to international migration flow data enables both a model selection based on the observed and missing information, and the appropriate variance-covariance matrix for parameter estimates in such a model, to be determined. The latter of these results will thus allow confidence intervals for fitted values, including the imputed missing data estimates, to be obtained.

6.2 Properties of the EM Algorithm

The derivation of the SEM algorithm is dependent on both analytical expressions for the rate of convergence of the EM algorithm and manipulations of the asymptotic variance-covariance matrix of parameter estimates. Both of these are further outlined in the following subsections.

6.2.1 Rate of Convergence in the EM Algorithm

For the EM algorithm described in Section 5.3, the mapping $\boldsymbol{\theta} \rightarrow M(\boldsymbol{\theta})$ from the parameter space of $\boldsymbol{\theta}$, to itself is implied. Consequently for every iteration,

$$\boldsymbol{\theta}^{r+1} = M(\boldsymbol{\theta}^{r}), \text{ for } r = 0, 1, \ldots. \tag{6.1}$$

Hence, when the parameters converge to a stationary point $\boldsymbol{\theta}^*$ and a given $M(\boldsymbol{\theta})$ is continuous,

$$\boldsymbol{\theta}^* = M(\boldsymbol{\theta}^*). \tag{6.2}$$

As Meng and Rubin (1991) noted, in the neighbourhood of $\boldsymbol{\theta}^*$ by a Taylor series expansion

$$\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^* \approx (\boldsymbol{\theta}^{r} - \boldsymbol{\theta}^*)\mathbf{DM}, \tag{6.3}$$

where

$$\mathbf{DM} = \left(\frac{\partial M_j(\boldsymbol{\theta})}{\partial \theta_i} \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{6.4}$$

is a $p \times p$ Jacobian matrix for $M(\boldsymbol{\theta}) = (M_1(\boldsymbol{\theta}), \ldots, M_p(\boldsymbol{\theta}))$, known as the rate matrix. Thus the EM algorithm converges in a linear fashion in the neighbourhood of $\boldsymbol{\theta}^*$. The rate of convergence is governed by the rate matrix which, as shown in the following subsection, represents the fraction of missing information.

6.2.2 Asymptotic Variance-Covariance Matrix

The distribution of the complete data, y, can be factored into components of observed data, y_o, and missing data, z:

$$f(y|\boldsymbol{\theta}) = f(y_o, z|\boldsymbol{\theta}) = f(y_o|\boldsymbol{\theta}) f(z|y_o, \boldsymbol{\theta}), \tag{6.5}$$

where $f(y_o|\boldsymbol{\theta})$ is the density of observed data and $f(z|y_o,\boldsymbol{\theta})$ is the density of missing data given the observed data. Thus the log likelihood of $\boldsymbol{\theta}$ given y is

$$l(\boldsymbol{\theta}|y) = l(\boldsymbol{\theta}|y_o) + \log f(z|y_o,\boldsymbol{\theta}). \tag{6.6}$$

When working with complete data it is common practice to use the asymptotic variance-covariance matrix, $\mathbf{V}$ of $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ based on y. This can be found using the inverse of the observed information matrix,

$$\mathbf{V} = \mathbf{I}^{-1}(\boldsymbol{\theta}^*|y), \tag{6.7}$$

where $\mathbf{I}(\boldsymbol{\theta}) = \frac{\partial^2 logf(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$. However, in the presence of missing data this function can be difficult to evaluate directly using methods based solely on the observed data. With incomplete data the observed component of the complete data information, $\mathbf{I}_{oc}$, can be deduced as

$$\mathbf{I}_{oc} = E[I_o(\boldsymbol{\theta}^*|y_o)|y_o,\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, \tag{6.8}$$

This can be obtained from the inverse of the variance-covariance matrix when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ using standard methods as in (6.7). In order to deduce the missing information consider the second derivatives of (6.6) averaged over $f(z|y_o,\boldsymbol{\theta})$ and evaluated for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ as

$$\mathbf{I}_o(\boldsymbol{\theta}^*|y_o) = \mathbf{I}_{oc} - \mathbf{I}_m, \tag{6.9}$$

where $\mathbf{I}_o = -E\left[\frac{\partial^2 logf(y_o|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|y,\boldsymbol{\theta}\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and $\mathbf{I}_m = -E\left[\frac{\partial^2 logf(z|y_o,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|y_o,\boldsymbol{\theta}\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$. $\mathbf{I}_m$ can be thought of as the missing information and thus (6.9) can be interpreted neatly as

$$\textit{observed information = complete information - missing information}, \tag{6.10}$$

otherwise known as the missing information principle of Orchard and Woodbury (1972). The above equation may also be written as

$$\mathbf{I}_o(\boldsymbol{\theta}^*|y_o) = (I - \mathbf{I}_m\mathbf{I}_{oc}^{-1})\mathbf{I}_{oc}, \tag{6.11}$$

where I is an identity matrix. As Dempster et al. (1977) noted, if $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^r)$, the augmented log likelihood of (5.1), is maximized in the M step by setting its first derivative to zero, then the differential of the parameter mappings in (6.4) is

$$\mathbf{DM} = \mathbf{I}_m\mathbf{I}_{oc}^{-1}. \tag{6.12}$$

This property can be substituted into (6.11) and inverted to give an expression for the asymptotic variance-covariance matrix of the parameter estimates from incomplete data,

$$\begin{aligned} \mathbf{V} &= \mathbf{I}_{oc}^{-1}(I - \mathbf{DM})^{-1} \\ &= \mathbf{I}_{oc}^{-1} + \Delta\mathbf{V} \end{aligned} \tag{6.13}$$

where $\Delta\mathbf{V} = \mathbf{I}_{oc}^{-1}\mathbf{DM}(I - \mathbf{DM})^{-1}$, and $\mathbf{V}$ is a symmetric matrix.

6.3 Supplemented EM algorithm

The estimation of $\Delta\mathbf{V}$ in (6.13) can be obtained using the SEM algorithm introduced by Meng and Rubin (1991). The SEM algorithm consists of three parts, the evaluation of $\mathbf{I}_{oc}^{-1}$, the computation of $\mathbf{DM}$ and the evaluation of $\mathbf{V}$. $\mathbf{I}_{oc}^{-1}$ can be obtained relatively easily using the standard complete data variance-covariance matrix evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Hence, $\mathbf{I}_{oc}^{-1}$ can be determined from the final imputations in the E step. Computations of the $\mathbf{DM}$ matrix are often more complicated. As described in the previous section, the $\mathbf{DM}$ matrix represents the differential of parameter mappings during the EM algorithm. Hence, each element of the matrix represents a component wise rate of convergence of iterations in the EM algorithm. This can be derived numerically by considering the (i, j)th element of the $\mathbf{DM}$ matrix to be a_{ij} and defining $\boldsymbol{\theta}^r(i)$ to be a semi-active parameter set:

$$\boldsymbol{\theta}^{(r)}(i) = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(r)}, \theta_{i+1}^*, \ldots, \theta_p^*), \tag{6.14}$$

where only the ith component in $\boldsymbol{\theta}^r(i)$ takes a value different from its maximum likelihood estimate. Thus from (6.4) we can define a_{ij} as

$$\begin{aligned} a_{ij} &= \frac{\partial M_j(\boldsymbol{\theta}^*)}{\partial \theta_i} \qquad (6.15)\\ &= \lim_{\theta_i \to \theta_i^*} \frac{M_j(\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i, \theta_{i+1}^*, \ldots, \theta_d^*) - M_j(\boldsymbol{\theta}^*)}{\theta_i - \theta_i^*} \\ &= \lim_{r \to \infty} \frac{M_j(\boldsymbol{\theta}^{(r)}(i)) - \theta_j^*}{\theta_i^{(r)} - \theta_i^*} \equiv \lim_{r \to \infty} a_{ij}^{(r)}. \end{aligned}$$

As $M(\boldsymbol{\theta})$ is obtained automatically by the output of the EM algorithm, all elements of a_{ij} can be estimated using a record of M-step iterations, including the converged set of estimates $\boldsymbol{\theta}^*$ (which could have been estimated from another procedure) and a set of starting points $\boldsymbol{\theta}^{(1)}$ not equal to $\boldsymbol{\theta}^*$ in any component. These are used in each cycle of the SEM algorithm consisting of three steps:

1. Run a single iteration of the EM algorithm given $\boldsymbol{\theta}^{(r)}$ to obtain $\boldsymbol{\theta}^{(r+1)}$.
   Repeat steps 2 and 3 for $i = 1, \ldots, p$.

2. Calculate a semi-active parameter set $\tilde{\boldsymbol{\theta}}^{(r)}(i)$ from (6.14) to be used as a current estimate of $\boldsymbol{\theta}$. Run a single iteration of the EM algorithm to obtain $\tilde{\boldsymbol{\theta}}^{(r+1)}(i)$.

3. Calculate the ratio

$$a_{ij}^r = \frac{\tilde{\theta}_j^{(r+1)}(i) - \theta_j^*}{\theta_i^{(r)} - \theta_i^*}, \text{ for } j = 1, \ldots, p. \tag{6.16}$$

After a single cycle, estimates of $\boldsymbol{\theta}^{(r+1)}$ and $\{a_{ij}^{(r)}, i, j = 1, \ldots, d\}$ are obtained. The algorithm repeats for r cycles until the sequence of $a_{ij}^{(r^*)}$, $a_{ij}^{(r^*+1)}$ is stable for some r. As different parameters in the initial $\boldsymbol{\theta}^{(r)}$ may be closer to $\boldsymbol{\theta}^*$ than others for any $\theta_i^{(r)}$, the number of iteration steps taken for stability of given elements of $\mathbf{DM}$ may vary. Hence, when all elements of the ith row of $\mathbf{DM}$ have been obtained, there is no need to repeat steps 2 and 3 for that parameter in subsequent iterations. Given the value of the converged $\mathbf{DM}$ matrix and $\mathbf{I}_{oc}^{-1}$ the asymptotic variance-covariance matrix for parameter estimates can be obtained using the expression of $\mathbf{V}$ in (6.13) where the resulting matrix should be numerically symmetric.

6.4 Akaike Information Criterion for Incomplete Data

Finding a suitable dimension for parameters $\boldsymbol{\theta}$ can be undertaken by comparing several models based on their values of an information criteria, such as the Akaike Information Criterion (AIC) of (5.3). This criterion can be thought of as a measure of separation between a fitted model for the incomplete data, $f(y_o|\widehat{\boldsymbol{\theta}})$ and the true or generating model which gave rise to the incomplete data, say $f(y_o|\boldsymbol{\theta}_g)$. Shimodaira (1994) noted that it may be more natural to use a criteria based on the complete data, assessing the separation between the fitted model $f(y|\widehat{\boldsymbol{\theta}})$ and the generating model $f(y|\boldsymbol{\theta}_g)$. Such an approach is advantageous for a number of reasons. Firstly, model families fitted by the EM algorithm are postulated for the complete data, and thus the model selection should reflect both observed and missing data. Secondly, as Meng and Rubin (1991) noted, the EM algorithm utilizes the computing power and complete data tools in handling missing data. Thus, the use of complete data tools can be incorporated to calculate a selection criteria based on these quantities rather than an analogous incomplete data criteria. Finally, as illustrated in (6.6) the complete data is a product of observed and missing densities. If the missing density is substantially affected by deviations of the true parameters then a model selection criteria based on incomplete observed data may not account for these alterations effectively.

In order to address these problems Cavanaugh and Shumway (1998) developed a criterion based on the complete data. The AIC of (5.3) when typically considered in an the complete data setting can be represented as

$$AIC = -2l(\boldsymbol{\theta}|y) + 2p, \tag{6.17}$$

where the first and second terms on the right hand side are commonly referred to as the goodness of fit and penalty terms respectively. When the observed data is incomplete Cavanaugh and Shumway (1998) derived a equivalent statistic to (6.17) for the complete data as

$$\begin{aligned} AICcd &= -2Q(\boldsymbol{\theta}|\boldsymbol{\theta}) + 2p + 2\ \mathrm{trace}[\mathbf{I}_{oc}(\boldsymbol{\theta}|y_o)\mathbf{I}_{oc}^{-1}(\boldsymbol{\theta}|y_o)\mathbf{DM}(I-\mathbf{DM})^{-1}] \\ &= -2Q(\boldsymbol{\theta}|\boldsymbol{\theta}) + 2p + 2\ \mathrm{trace}(\mathbf{DM}(I-\mathbf{DM})^{-1}), \end{aligned} \tag{6.18}$$

where the goodness of fit in the first term is twice the augmented log likelihood of (5.1). The penalty term is formed by the summation of the AIC penalty term and $\mathbf{I}_{oc}^{-1}(\boldsymbol{\theta}|y_o)\mathbf{DM}(I-\mathbf{DM})^{-1}$. As seen in (6.13), the latter term represents the increase in variance of $\boldsymbol{\theta}$ due to missing information when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Thus the trace of the penalty term in (6.18) can be conveniently viewed as a measure on the amount of data which is missing in y, or more precisely as a measure of the extent to which the missing data affects the fitted model. It is useful to note that if there was no missing data the trace of this term would be zero (as $\mathbf{DM} = 0$) and also, as the amount of missing data increases, so will the penalty term.

6.5 Estimates of Precision for Missing Data in International Migration Flow Tables

The SEM algorithm can be utilized in the estimation of international migration flow tables. In the remainder of this chapter, the analysis of the algorithms application to the EU15 data (where the harmonized data were treated as observed values) is divided into two parts. First, the convergence of the a_{ij} elements is studied in order to obtain a better understanding of the SEM algorithm for which a careful monitoring is required when considering large parameter vectors. Second, the AICcd statistics is used to select a model using the complete data, rather than the incomplete observed data (as used in the previous chapter). This is undertaken using a systematic fitting of a range of main effects models.

6.5.1 Convergence

In order to calculate the asymptotic variance-covariance matrix for parameter estimates, three functions were written in S-Plus for

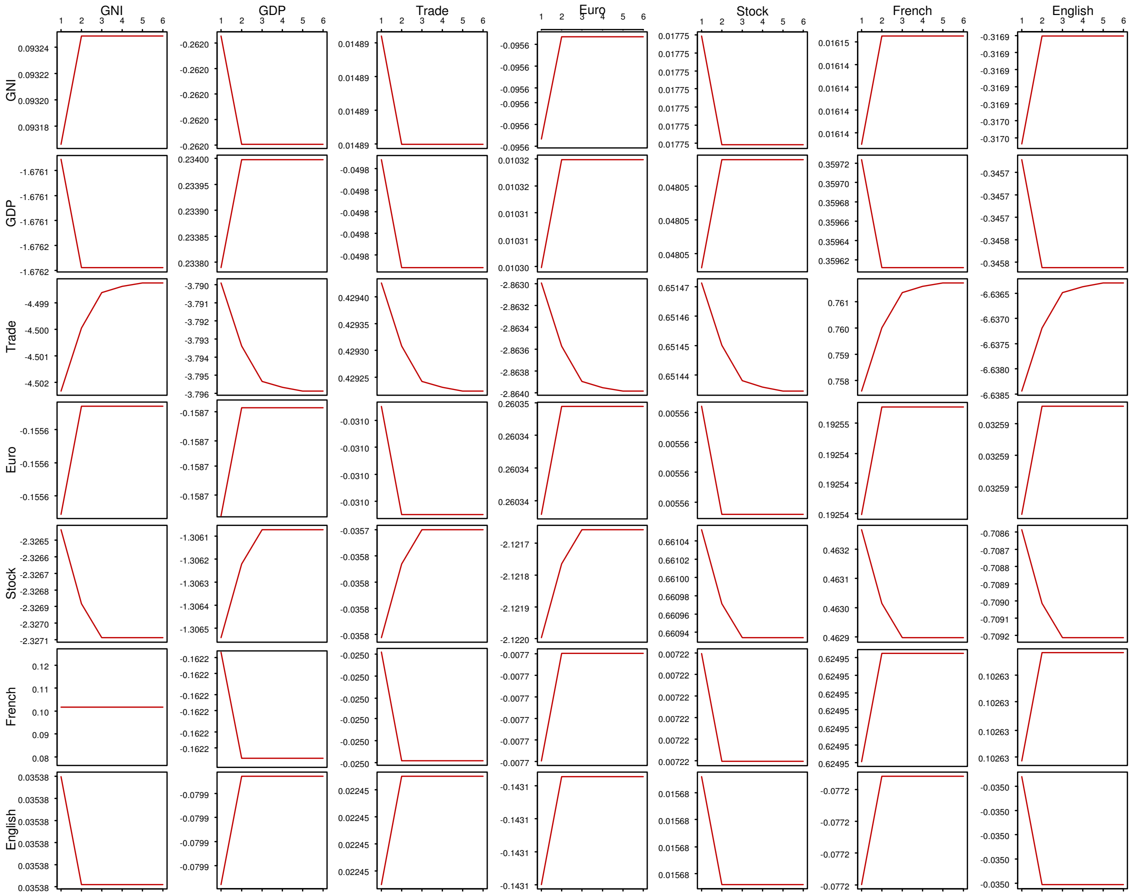
(a) the calculation of a single step in the EM algorithm,

(b) the numerical calculation of a Jacobian matrix of model parameters,

(c) the SEM algorithm.

These functions are displayed in the Appendix. The first function (`em`) is a general routine for a single step of the EM algorithm to be called in other functions. This requires values of the initial set of parameters, a `negbin` model from the MASS library of S-Plus Venables and Ripley (2003) and a data frame. This function can be run inside a loop routine until the difference in consecutive parameter values (`beta`) or the augmented log likelihood (`m$twologlik`) given by the function are less than a desired stopping criteria. A routine for the numerical calculation of the Jacobian matrix (`jac`) was used to cacluate the **DM** matrix from (6.16) for a given initial set of parameters (`b.init`) and the maximum likelihood estimates (`b.star`). The SEM algorithm is implemented using the `sem` function. Initially two consecutive Jacobian matrices from a given set of initial parameter values are calculated in order to deduce all elements of a two initial rate matrices $a_{ij}^{(1)}$ and $a_{ij}^{(2)}$. In further iterations of a_{ij}, calculations are performed within a loop in the `sem` function, calculating elements for rows where the maximum difference in elements are above the stopping criteria. Excluding further calculations in rows where all elements have already converged is beneficial from two standpoints. First, computing speed is increased, which is particulary relevant for the spatial interactions models with large numbers of

Figure 6.1: Trace of **DM** Matrix for Selected Main Effects Model Parameters



parameters used. Second, without a row-dependent stopping criteria, previously converged components of the **DM** matrix can become unstable and hence convergence would not obtained.

A trace of iterations in the **DM** matrix is displayed in Figure 6.1 for the main effects model found in Section 5.4 with a tolerance level of 10^{-3}. Excluded are row and columns for the dispersion, intercept, and origin and destination terms. The traces demonstrate how for the selected parameters, convergence of elements in the **DM** matrix is dependent on the row, whereby some rows converge quickly to a stable values whilst others, such as trade and stock, take longer. For the element of **DM** matrix representing the covariance between French and GNI parameters only a very small change occurred before convergence at the third iteration, which could not be illustrated effectively by graphics in S-Plus 6.2. Using the converged values of the **DM** matrix an estimated asymptotic variance-covariance matrix for parameter estimates was obtained using (6.13). The lower right hand corner of the final matrix corresponding to the main effects parameters is

$$\begin{pmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & 0.12811 & -0.02132 & 0.00329 & -0.00153 & -0.00253 & 0.00696 & 0.00511 \\
\cdots & -0.02051 & 0.18688 & -0.00067 & 0.00083 & 0.00037 & -0.00676 & -0.00102 \\
\cdots & 0.00353 & -0.00080 & 0.00148 & -0.00031 & -0.00053 & -0.00335 & -0.00535 \\
\cdots & -0.00130 & 0.00061 & -0.00032 & 0.01184 & -0.00054 & 0.00187 & -0.01236 \\
\cdots & -0.00274 & 0.00046 & -0.00051 & -0.00055 & 0.00064 & -0.00073 & 0.00133 \\
\cdots & 0.00706 & -0.00635 & -0.00353 & 0.00176 & -0.00059 & 0.06145 & 0.01853 \\
\cdots & 0.00344 & -0.00025 & -0.00544 & -0.01187 & 0.00143 & 0.01811 & 0.13682
\end{pmatrix},$$

(where parameters are arranged in the order given in Figure 6.1). The square of diagonal elements represents the variance of parameter estimates. These can be compared to the standard errors for the main effects model in Table 5.1.

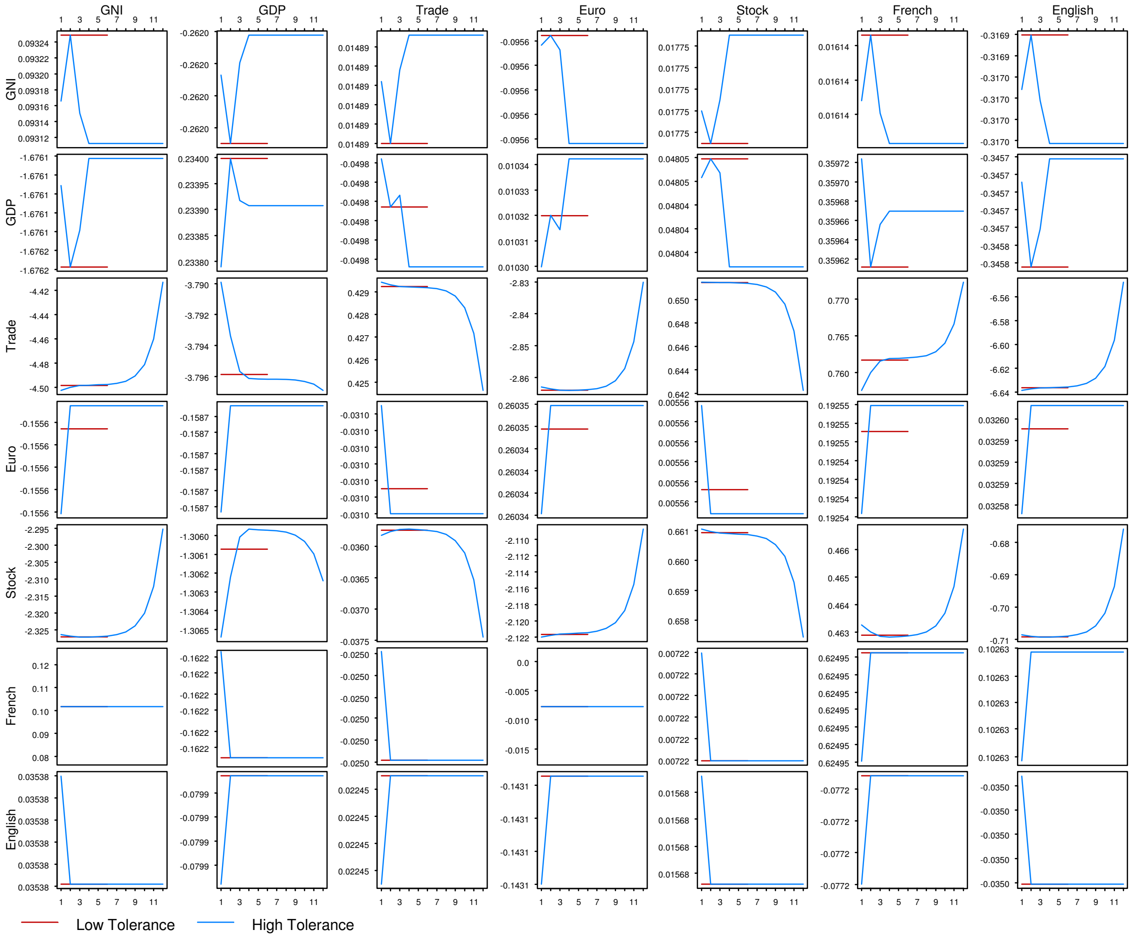
This matrix is symmetric when rounded to two decimal places. A more precise measure of the variance-covariance matrix could not be obtained. A tolerance level of 10^{-10} was taken for an estimate of `b.star` from the EM algorithm. Meng and Rubin (1991) suggest that the square root of the EM algorithms stopping criteria should be used (i.e. 10^{-5}). However, with such a tolerance level the convergence for all elements of the **DM** matrix is never obtained.

This feature is displayed in Figure 6.2 where the traces from Figure 6.1 are also plotted, but on the larger vertical axes appear flat. For most rows, convergence takes longer under a higher tolerance level, and for GNI, GDP and the Euro parameters slightly different values are obtained. Rows in the **DM** matrices for trade and stock parameters never converge. Only the first 12 iterations are shown in Figure 6.2 but the divergence of trade and stock parameters values continues until the number of iterations equals that of the EM algorithm. This ultimately creates an asymmetry in the respective rows and columns of the asymptotic variance-covariance matrix. This failure may be due to the large size of the **DM** matrix in comparison to examples used by Meng and Rubin (1991). Most plots in rows where convergence is not obtained have some degree of flatness in early estimates and hence convergence of their individual elements may have been obtained using a less stringent tolerance level. However, new values for a complete row are estimated if any element in the selected row fall short of the stopping criteria and thus elements that may have appeared stable continue to be estimated. An example of this process is illustrated in selected plots in Figure 6.2. For the **DM** elements related to the covariance of stock and GDP and the variance of stock, consecutive estimates do not stabilize below the stopping criteria, shown by the traces becoming nearly horizontal, but not completely flat, unlike other elements in the same row. With a lower tolerance level, the algorithm would have stopped estimating elements in this row when the troubled elements were nearly horizontal.

6.5.2 Modelling of Complete Data

As no implementable stepwise model selection routine existed for incomplete data, a fit all models function was written to run the SEM algorithm on the complete range of main

Figure 6.2: Trace of **DM** Matrix for Selected Main Effects Model Parameters for Low (10^{-3}) and High (10^{-5}) Tolerances



effects models from the covariate set proposed in Section 5.4.1. Included as a prerequisite in all models were origin and destination covariates. Consequently, from the 12 possible parameters (including time), there existed $\sum_{p=0}^{12} \frac{12!}{p!(12-p)!} = 4096$ different models. For each of these models the EM algorithm was run to obtain estimates for `b.star` in the `sem` algorithm. Converged estimates of the **DM** matrix from the `sem` algorithm were then used to calculate the AICcd statistic of (6.18). This was performed with a stopping criteria of 10^{-10} for the EM algorithm and 10^{-3} for the SEM algorithm.

Table 6.1: AIC, AICcd and Number of Parameters (p) for Selected Models

Selection	AIC	AICcd	p
`stepAIC`	12101.26	15366.02	36
Minimum AICcd	12102.10	15363.04	38
Minimum AIC	12098.02	15365.23	38

The model found with the lowest AICcd included the same covariates of the model found by the `stepAIC` function in the previous chapter (GNI, GDP the Euro currency area, trade, migrant stocks and the level of French and English), as well as, time and distance parameters. The value of the AICcd statistic, shown in Table 6.1 is higher than the AIC for the same model due to the expansion of the penalty term in (6.18). Imputation for model with the lowest AICcd were very similar to that from the original main effects model in Table 5.1 as the exponentiated estimates for time and distance were both near unity (0.9808 and 1.0699, respectively) whilst values for other parameters altered only slightly. For comparative purposes, the AIC for the observed data was also found for each model. The model with smallest AIC included parameters for time and population (again with exponentiated values close to unity 0.9630 and 1.0789, respectively) in addition to the main effects model in Table 5.1. The model selected using the `stepAIC` function had the tenth smallest AIC of all possible models. The number of parameters in the model with the smallest AIC is equal to that of the model with the smallest AICcd. Cavanaugh and Shumway (1998) found in a simulation study on models for bivariate normal data that this result is reasonable, noting that the AICcd tended to overfit (select more parameters than the true model) to a comparable or to a slightly lesser degree than the AIC. This property was attributed to be a result of incorporating the missing data into the penalization term, lacking in the AIC statistics.

The estimated asymptotic variance-covariance matrix of the parameter estimates from the SEM algorithm can be used to create measures of precision for a vector of imputations, $\mathbf{z}$. These can be expressed as confidence intervals, where

$$\begin{aligned} \mathrm{Var}(\log \mathbf{z}) &= \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) \qquad (6.19) \\ &= \mathrm{Var}(\mathbf{X}\mathbf{V}\mathbf{X}^{\mathbf{T}}) \end{aligned}$$

Hence, scaling covariate values in the model matrix by the estimated asymptotic standard errors and a Z-value based on a 95% confidence level imputation is

$$\log \mathbf{z} \pm 1.96\mathbf{X}\widehat{\mathbf{V}}\mathbf{X}^{\mathbf{T}}, \qquad (6.20)$$

where the logarithmic transformation is applied component wise. Exponentiated confidence limits are shown for imputations given under the model selected by the AICcd in Figure 6.3 for the EU15 flows. To allow a clearer illustration, flows between the six countries of the European Coal and Steel Community (ECSC), a forerunner of the EU are shown in Figure 6.4. The width of intervals in these plots is greater for larger flows. These bounds demonstrate that with a 95% confidence level the flow value under the main effects model (in Table 5.1) lies within this interval. Note, these intervals only represent the variability of the mean response, derived from the parameter estimates. An additional term is required to fully represent the variability of the predicted flows.

6.6 Summary and Discussion

The SEM algorithm provides a useful technique when applied to international migration flow tables, where data is often incomplete. Obtaining an estimate of the asymptotic

Figure 6.3: Imputations and 95% Confidence Bounds of Estimated Migration Flows (000's) from each Origin-Destination Combination of EU15, 2002-2006.

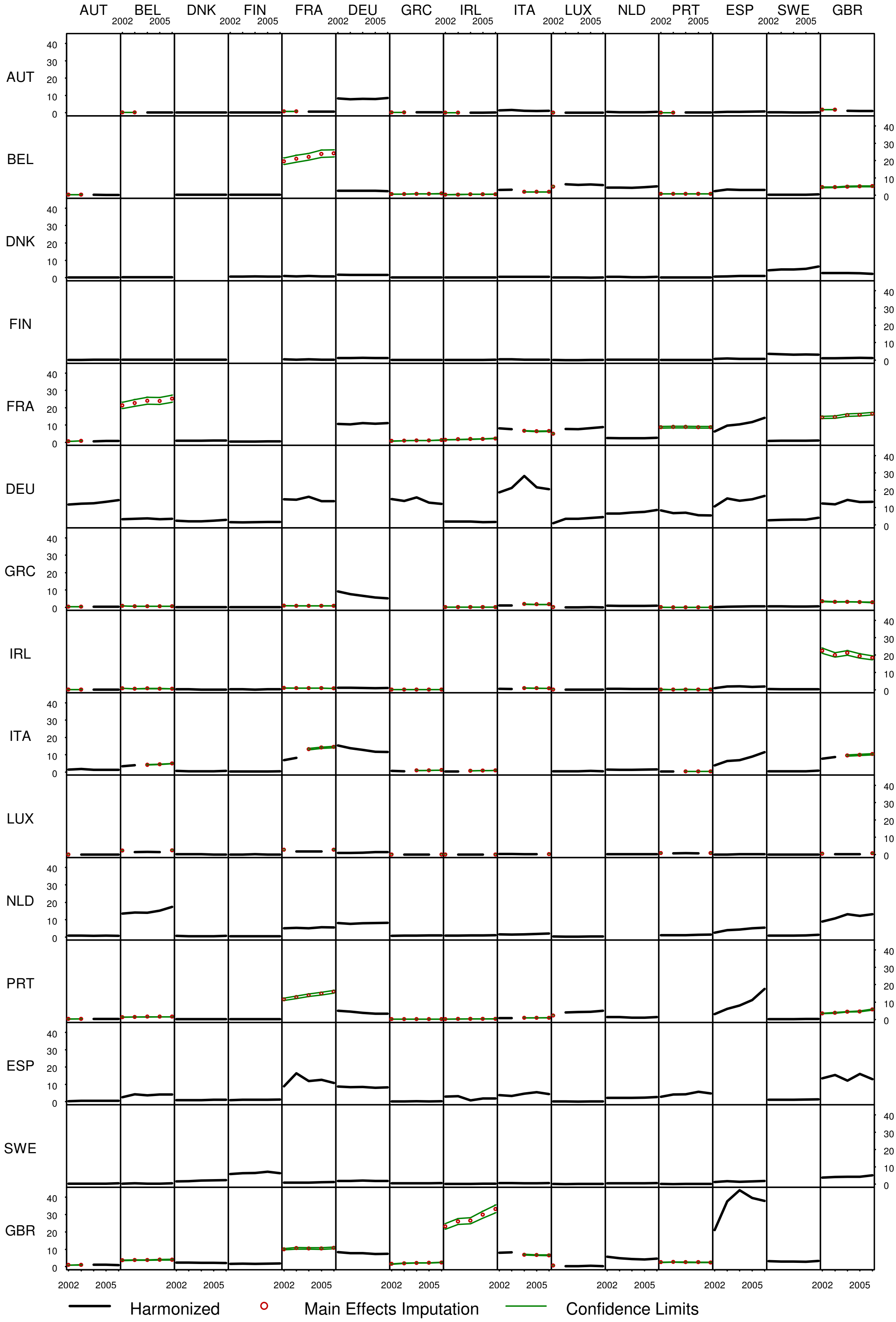
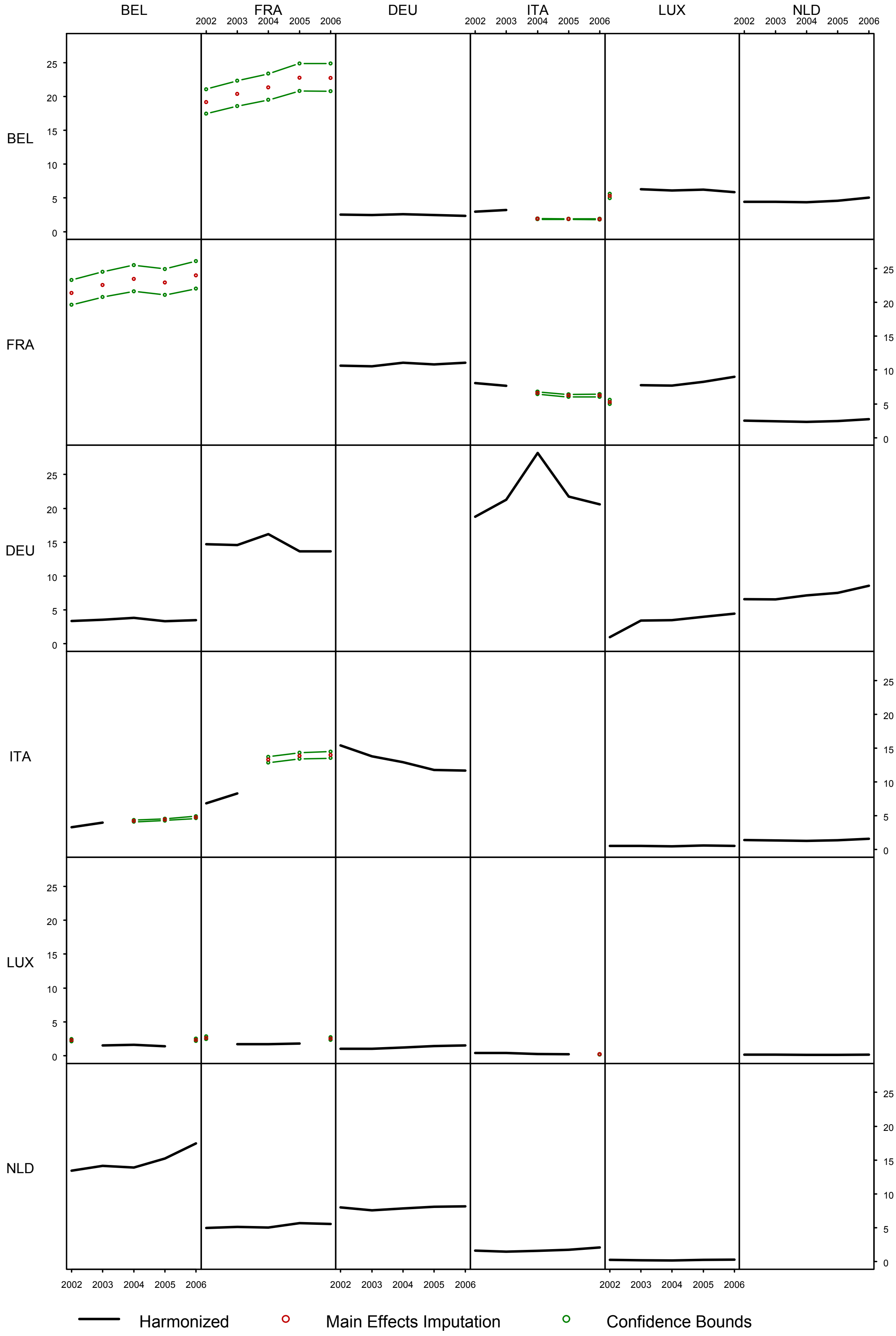


Figure 6.4: Imputations and 95% Confidence Bounds of Estimated Migration Flows (000's) from each Origin-Destination Combination of ECSC, 2002-2006.



variance-covariance is beneficial in gaining a better understanding of the variance of imputed values, allowing a confidence interval to be constructed. In addition, the SEM algorithm estimates the rate of convergence matrix, **DM**, which is used in the AICcd statistics to select models based on the complete data.

The SEM algorithm can be performed using only the code for an EM algorithm, computations for asymptotic complete data variance-covariance matrix and standard matrix procedures. More complicated methods for estimation of asymptotic variance-covariance matrices such as that of Louis (1982), require the observed information to be approximated using conditional expectations of first and second moments of the gradient and curvature of the complete data introduced within the EM framework. However, as Meng and Rubin (1991) noted this method (along with others such as Carlin (1987) and Meilijson (1989)), besides requiring evaluation of the likelihood, are subject to the inaccuracies and difficulties of any numerical differentiation procedure with large matrices. The SEM algorithm is also more stable than alternative methods which rely on pure numerical differentiation. In the SEM algorithm, the rate of change matrix is being added to an analytically obtained matrix ($\mathbf{I}_{oc}$), rather than the whole covariance matrix. This allows a degree of stability, as when missing data is plentiful, the convergence of EM algorithm is slow and hence a long sequence of iterates is provided from the linear rate of convergence leading to high levels of accuracy. When there is less missing data, the convergence of the algorithm is quick but the estimate of $\mathbf{I}_{oc}$ is fairly accurate. Hence, the increase in variance from missing data does not dominate the calculation the complete variance-covariance matrix.

The stability of the SEM algorithm and the ease of implementation were exploited in this chapter by fitting all possible model formulations when choosing from 12 different parameters. This resulted in a model fairly similar to that obtained using the observed data. This is predominantly explained by the automatic inclusion in all models of the origin and destination terms. These provide a lot of information on the push and pull effects for each country, which limits the other parameters to modify the interaction between countries, conditional on the inclusion of the country specific variables. Alternatively these terms could be excluded and hence a gravity model formulation for international migration flow tables would be used. However, as Flowerdew and Lovett (1988) and Flowerdew (1991) noted for internal migration tables, these often provide worse fits. Another alternative would be to search amongst all interaction models for the lowest AICcd. With a 15 × 15 table this would lead to a vast amount of covariates with parameters for multiple levels to be estimated. As seen in the previous chapter with such models parameter identification with only five time periods becomes an issue. Further consideration could be taken for selecting models based on the AICcd for interactions between countries with incomplete data and unsatisfactory imputations from a main effects model (as performed in the last chapter using the AIC for selecting interactions with France as an origin and destination). A routine to fit all models with the 22 extra covariates would involve fitting approximately 6.87×10^9 potential models using the SEM algorithm. A suitable model could be more efficiently found using a stepwise search algorithm such as `stepAIC` adapted to select models based on the AICcd.

The stopping criteria for the SEM algorithm, when fitting spatial interaction models, were taken to be lower than that recommended. This was necessary for convergence and, in some cases, resulted in different values in the **DM** matrix than under more strict stopping criteria. When such differences occurred, they were very small and did not effect the symmetry of the variance-covariance matrix in their respective rows and columns. For some parameters, their elements in the **DM** did not converge for high tolerance levels. Parameter estimates are from a model distribution not in the exponential family and thus the estimate of the dispersion parameter depended on asymptotic approximations in M-Step, using the `glm.nb` function. Consequently, estimates are based on linearizations using the Newton-Raphson routine, which as noted may create numerical inaccuracies in comparison to parameter estimates of $\mathbf{V}$, from distributions in the exponential family, using for example, IRLS. These inaccuracies may have affected the calculation a_{ij} in later iterations, which after a certain amount of iterations begin to use ever smaller numbers in both the numerator and denominator leading to the divergent behaviour shown in Figure 6.2.

In conclusion, the SEM algorithm is a powerful tool when modelling international migration flow tables. It allows information on the second moment to be derived from the complete data asymptotic variance-covariance matrix for parameter estimates and for the selection of a model to account for missing data. This facilitates the creation of a confidence interval for model based imputations and thus provides added information to data users to gain a better understanding of the reliability of estimated flows where no previous data exist.

Chapter 7

Conclusion

7.1 Summary

This study applied computationally intensive mathematical and statistical techniques to develop a methodology to estimate international migration flow tables of comparable data. Such tables commonly suffer from problems of inconsistent and incomplete data which previous estimation frameworks outlined in Chapter 3 failed to fully address.

The methodology developed in this thesis can be judged against the desirable criteria for estimating international migration flow tables introduced in Table 1.1. Complete estimates were obtained in Chapter 5 by modelling incomplete migration flow tables. Parameters for these models were estimated using the EM algorithm which also provided imputations for unknown migration flow counts. Consistent migration flows across multiple nations were obtained using estimated correction factors to scale reported data. In Chapter 4, constrained optimization techniques were used to estimate correction factors alongside expert opinion on the quality of migration statistics produced by national statistics institutes. The calculations of these correction factors required that reported flows were of a reasonable quality and hence the scaling of reported data was only performed for data from reliable sources. Checks for reliability were made throughout each stage of the methodology. These were partly undertaken by considering data across multiple time periods, which is discussed further in this chapter. Reliability checks for inconsistent data were made by comparing observed distributions for reported data from reliable sources with estimates. As discussed in Chapter 4, for receiving data these distributions remain unchanged as the estimates were a scaled version of reliable reported data. Reliability checks for missing data considered estimates in relation to expected results under international migration theory. As demonstrated, in Chapter 5 models were expanded to include further covariates to help improve estimates of flows to and from France which were initially believed to be unreliable. In Chapter 6, measures of precision for missing data were derived by estimating the asymptotic variance-covariance matrix of parameter estimates using the SEM algorithm. Combined, the procedures for dealing with inconsistencies and incompleteness, introduced in this thesis, allowed the estimation of migration flow tables that are comparable across nations and time.

The methods in this thesis used a number of the desirable properties for estimation techniques introduced in Table 1.1. In Chapter 5, Missing data estimates were based on models, selected from fits on the observed data, for which parameters were estimated using the EM algorithm. In Chapter 6, the AICcd was used to select a main effect model based on the complete data. Model based techniques for imputations allowed a great deal of flexibility to ensure missing data are of reasonable quality. For example, complex models were fitted in Chapter 5 that included interaction terms to allow more realistic imputations to be estimated. However, the use of multiple interactions caused issues with parameter identification due to the limited amount of observed data. Expert opinion was used in estimating consistent and complete flows. This included selecting data to be unchanged, scaled (or ignored) in the constrained optimization procedure and the collection of appropriate covariate factors for model based imputations. The latter of these may be beneficial when the harmonized data are heavily reliant on the selected distance measure. Although the distance measures studied in Chapter 4 produced correction factors that were alike in most time periods, for different data the choice of distance measure might be very influential on the estimated harmonized data. In such a case, an alternative model is likely to be selected for the imputation of missing flows by a stepwise routine. However, expert opinion can help inform the selection process if the resulting estimates for missing cells are judged to be unrealistic. As discussed, this might involve the addition of new covariates or interaction terms to help improve model fits.

The methodology presented in this thesis can be relatively easily replicated in S-Plus/R given the data and the functions supplied in the Appendix. The constrained optimization techniques and modelling of incomplete migration flow tables using the EM algorithm can be applied to models for alternative international migration tables. This can include both smaller or larger tables and additional data for previous or subsequent time periods. In this thesis, EU15 nations were used to enable effective models for flows between politically similar countries with only a few main effects. In a more diverse set of nations (or a longer time period) additional care would be required to obtain more reliable estimates. This might be in the form of more correction factors to account for changes in data sources, or additional covariates to account for more diverse sets of nations.

The remainder of the current chapter summarizes some of the key results of this thesis. These will be discussed alongside some of the selected contributions found in this thesis and recommendations for potential future research. This will be broken into five areas: estimating tables over time, accounting for counts of known migrants with unknown origin and destinations, ignoring poor quality data, model selection and variation measures. A more general discussion on the conclusions from this thesis is then put into the context of international migration estimation from the modelling and data perspectives.

7.1.1 Estimation Over Time

The relative stability in migration definitions and data collection systems provides a basis for harmonizing international migration flow data. These can be visualized through plots of selected flows, as demonstrated by Kupiszewska and Nowok (2008) or through plots of migration flow tables over time as shown in Figure 4.1. Previous methods for harmonization of reported data used differing methods and typically concentrated on tables for a single year (with the exception of Raymer and Abel (2008)).

In Chapter 3 of this thesis two existing frameworks (which incorporated systems for estimating missing data as well) were outlined. The framework of Poulain (1993) used a constrained optimization to minimize a single distance function of scaled flow data, whilst the methodology of Raymer (2007) relied upon the demographic accounting equations in each nation. The latter of these methods proved difficult to evaluate due to manipulations in marginal estimates and interpolation methods for missing data. The former of the previous frameworks provided a useful basis for further analysis. Chapter 4 of this thesis explored different measures and alternative constrained optimization techniques. Comparisons of estimates were undertaken through evaluations based on the variance within the set of correction factors across time. Plots of estimates also allowed an easy comparison of different constraint systems. The most stable distance function was generalized over time to allow single correction factors for each data source to be estimated, under the assumption that definitions and data collection systems were unchanged.

Modelling only a single flow table could potentially restrict the number of model parameters to be identified, especially when data are incomplete. Including multiple tables for analysis and controlling for time allows a far greater number of parameters to be estimated. In addition, flows for which only partial data were available provided useful information in the estimation of parameters and comparison of imputations for flows where no data were present. Within the modelling framework outlined, such imputations could be further improved by controlling for specific origin-destination combinations that are partially observed by including the relevant dummy covariate in a potential model.

The benefits of expanding migration flow tables over time could be improved by using a longer series of migration data if available. Reported flows between the EU15 previous to 2002, provided by Eurostat, appeared incorrect. In this data, the presented sending data appeared to be reported by destinations (forming vertical patterns when arranged into a migration table) rather than origins. If such values were corrected, a greater amount of information could be used in the estimation of correction factors and imputations, given the assumptions of constant methods of migration data collection and definitions hold. If changes did occur in the data collection or definitions for a given country, additional parameters for before and after any structural break can be included in the estimation of the correction factor in place of a single parameter for the entire time period. Such modifications are easy to implement in the non-linear optimization routines outlined in Chapter 4. Further data across time may also help inform judgment of experts on the quality of data sources and inform the eligibility criteria for the estimation of harmonized values. From a modeling perspective longer time series could be alternatively handled using marginal models which may allow more complex categorical covariates to be fitted. Imputations for missing data under such models would require more intricate parameter estimation techniques in the M-step of the EM algorithm and an assumption for the correlation structure of data.

7.1.2 Accounting for Data Dissemination Problems

As a prelude to the estimation of correction factors, counts of known migrants with unknown origins or destinations were accounted for by distributing these flows according to the existing distributional patterns. For some countries, the addition of these values altered the reported flows greatly. Consequently distance measures for the harmonization process were modified. Previous constrained optimization procedures for migration data had not considered such values. Comparisons of these values across time in Figure 4.3 revealed some notable insights. For Spanish data the number of known migrants with unknown origin and destinations were extremely different in 2002 than in subsequent years. However, most counts to specific origin and destinations are fairly stable over the time period once the unknowns were accounted for, and the literature considered (Breem and Thierry, 2006b) suggested that changes to the data measurement occurred previous to the studied time period.

More widespread documentation of the unknown counts in international migration flow data and the use of expert opinion could help account for the allocation of these flows. For example, if large portions of the unknown counts were to or from countries in a different continent the assumption of an equal distribution should be altered to reflect this failure.

7.1.3 Ignoring Poor Quality Data

Careful consideration was taken in deciding the eligibility of countries for the estimation of correction factors to scale reported data. This decision was based on recent literature by Erf (2007) that gave a quantifiable comparisons between migration sources. As a result, data which were judged to be of poor quality are ignored to enable a more effective estimation of parameters. Replacement values for ignored data were provided by imputations from a spatial interaction model estimated using the EM algorithm. The ratings of Erf (2007) were also used to select data sources, for which correction factors would be constrained to be one, and hence act as a reference for all estimates. Before correction factors were estimated, sending data from countries with migration data exchanges were also ignored as they are repetitions of data collected by receiving partner countries.

Further research into the comparison of migration definitions and data collection techniques may further inform the decision to ignore lesser rated data sources. For example, ratings for sending and receiving data were treated equally although literature suggests that this is not the case. As comparable ratings are only provided within data types, no distinction could be made between sending and receiving data qualities. Ratings provided across all data sources may enhance the decisions for which data sources should be ignored, require a correction factor to be estimated or constrained. They may help inform a preference system to obtain a single flow value in each cell, where for example, a particular sending data source might be considered better than any other receiving data.

Alterations in the eligibility of data sources for the application of correction factors due to new expert opinions or further documentation can be easily incorporated. The use of non-linear optimization routines in statistical software allowed a great deal of flexibility to change constraints and use alternative distance measures. In addition, more realistic bounds for correction factors could be introduced. For example, tighter bounds in the parameter space could force estimates to be no lower or greater than a value supplied by expert opinion. Routines might also be easily constrained to harmonize data to an alternative set of countries' reported flows, which may use different timing criteria in their migration definition, such as a six month definition as used by multiple migration data sources in the EU15. Models might then be fitted to the new harmonized level of data using the EM algorithm to provide comparable data for shorter timing criteria.

Additional data on sending flows between countries that currently have data exchange agreements would enable more measures of data discrepancies to be obtained. For example, reported sending counts of movements from Denmark to other Nordic nations, which may already be collected but not published, would be valuable in estimating correction factors for all concerned countries. The inclusion of extra migration flow data from countries with reliable sources but outside the migration table of study could also be used to provide more distance measures in the estimation of parameter values. For example, receiving data from Norway is regarded to be of good quality and uses a one year definition (Erf, 2007). A distance measure between its estimates and other countries sending data may further improve the credibility of correction factor estimates.

7.1.4 Model Selection

The EM algorithm was used to impute missing migration flow values. An underlying negative binomial regression model in Chapter 5 was selected using a stepwise search routine to compare the AIC of models. This routine was initially run to select main effects parameters only, followed by a wider consideration for interaction terms. Although computationally fast this procedure was based on observed data, and hence made no consideration for the missing data. In addition, when parameters were fitted by implementing the EM algorithm problems occurred with identification for some levels of interaction covariates. This was due to the limited amount of observed data being used to estimate a large number of parameters. In Chapter 6, new main effects models were selected based on the complete data through comparisons of the AICcd. This required models to be fitted by implementing the SEM algorithm, slowing the computational time. No implementable stepwise routine existed to compare models based on the AICcd and hence an all models routine was used.

Interaction terms to improve imputations can be added by considering expert opinion. In Chapter 5, for flows to and from France considered interactions of origin and destinations with other covariates. Further improvements to the model fit, and hence imputations, could be undertaken by including other country specific interactions where recommended from data experts. Additional main effects and redefining the origin-destination relationships in existing covariates may also improve a models fit if selected. For example, comparative measures of unemployment or climate could be utilized if comparative measures for the duration of the time period studied 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 data collection process. Analysis of lagged or quadratic relationships may also provide useful contributions to models. Negative binomial regression models were used throughout the modelling process in this study. This was undertaken to account for the overdispersion in aggregate level migration data.

The selection of a main effects model based on the AICcd required a far greater number of calculations, and hence computational time than the stepwise model selection routine. A suitable model could be found more efficiently using a stepwise search algorithm adapted to select models based on the AICcd.

7.1.5 Measures of Variation

The SEM algorithm was used in Chapter 6 to obtain an estimate for the asymptotic variance-covariance matrix for parameter estimates, using only the code for an EM algorithm, computations for asymptotic complete data variance-covariance matrix and standard matrix procedures. This allowed a better understanding of the possible variation of imputed values under a selected model, allowing a confidence interval to be constructed. The stopping criteria for the SEM algorithm, when fitting spatial interaction models, were taken to be lower than those recommended. This was necessary for convergence as the dispersion parameter in the negative binomial distribution depended on a Newton-Raphson routine which created numerical inaccuracies in comparison to parameter estimates of distributions in the exponential family.

More accurate measures of the asymptotic variance for a selected model could be derived using alternative methods, such as Louis (1982), although the generalisability is more limited than the SEM, whereby conditional expectations of first and second derivatives of the complete data are required. This would prove problematic if fitting multiple models with different numbers of parameters.

Imputations and their confidence intervals assume that there exists no error in the estimation of correction factors. As shown in Figure 4.5 estimates are not constant across time, where some correction factors fluctuate greatly. In such situations, the assumption that a distance measure for the discordance between data collection and definitions from reliable data sources are fixed, may not be valid. Methods exist in the Bayesian paradigm (see the next section) that may allow this assumption to be relaxed and hence measures of variation to be more fully obtained for imputed values.

7.2 Context of Study

7.2.1 Modeling International Migration

There exists a wide range of literature on modeling migration (see for example, Massey et al. (1993) or Greenwood and Hunt (2003)). Due to data limitations, most empirical studies concentrate on internal migration or flows in or out of single countries. The use of migration flow tables allows comparisons of data sources to be analyzed and differences to be addressed. In this thesis, comparisons were extended over time to further analyze, correct for inconsistencies and enable the estimation of complex models for incomplete international migration tables. Such procedures relied on modern computationally intensive mathematical and statistical methodologies.

Alternative statistical approaches to the modelling of international migration data have been undertaken in a Bayesian framework. Brierley et al. (2008) proposed one such method using similar model component methodology of Raymer (2007) to estimate posterior distributions of both internal and international migration flows. For international data, prior distributions were assigned to parameters in a model similar to (3.11) under the assumption that receiving data took a log-normal distribution, allowing posterior distribution for a complete migration flow table to be obtained. As with the Raymer (2007) their analysis of Northern European receiving flow data produced final estimates for Lithuania which were too high and altered patterns in original good data such as Sweden. These problems had been driven by the assumption that all marginal data were complete and consistent. As discussed in Chapter 3, this is not the case with international migration where problems of both inconsistencies and incompleteness also appear in the marginal totals. In addition, simple models with only a single parameter to explain spatial interactions were used.

A Bayesian modelling framework for international migration flow tables could provide a number of advantages. Using a similar approach to the methodology outlined in this thesis we may express the distribution of migration flow data as observations from a true negative binomial distribution $y_{ijt} \sim NB(\mu_{ijt}, \alpha)$ where μ_{ijt} and α are the mean and dispersion parameters, respectively. Observations from this distribution, y_{ijtk} in receiving and sending countries are subject to a scaling dependent on the data source,

$$y_{ijt1}|r_j, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T \sim NB(r_j\mu_{ijt}, \alpha) \tag{7.1}$$

$$y_{ijt2}|s_i, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T \sim NB(s_i\mu_{ijt}, \alpha), \tag{7.2}$$

where $\log\mu_{ijt} = \mathbf{x}_i^T\boldsymbol{\beta}$. Hence, if individual level covariates exist in $\mathbf{x}_i^T$, the true model can be modified to set $\alpha = 0$, and thus a Poisson distribution is derived. Appropriate prior distributions for the parameters, $p(r_j)$, $p(s_i)$, $p(\alpha)$ and $p(\boldsymbol{\beta})$, where the dimension of $\boldsymbol{\beta}$ is already known, can also be expressed. This allows the joint posterior distribution for all parameters,

$$p(s_i, r_j, \alpha, \boldsymbol{\beta}|y_{ijt1}, y_{ijt2}, \mathbf{x}_i^T) = p(r_j)p(s_i)p(\alpha)p(\boldsymbol{\beta})p(y_{ijt1}|r_j, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T)p(y_{ijt2}|s_i, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T), \tag{7.3}$$

to be estimated. This can be computed using Markov Chain Monte Carlo methods, given some crude initial estimates for parameters; see for example Gelman et al. (2003, p283-307). Once obtained, estimates for the entire distributions of flows in each cell of a series of migration tables can be deduced.

Such an approach provides a number of advantages. As discussed previously, correction factor estimates may posses some element of error. A Bayesian model can provide a more realistic account for fluctuations in their estimates shown in Figure 4.5. As a result, variation in estimates of r_j and s_i can be accounted for in the estimation of the marginal distributions of $\boldsymbol{\beta}$, and thus the imputations for missing data. If an analyst assumes that there is no error in the difference between reported values, as was taken in this study, prior distributions for $p(r_j)$ and $p(s_i)$ may be defined with very low or zero variances. In the latter case, this would allow reported values from countries with excellent data collection methods and using the desired definition to be preserved.

7.2.2 International Migration Data

International migration flow data is often incomparable across multiple nations. The increasing concern of governments in the production of international migration statistics may in future lead to data provided by statistics institutes becoming more readily available and of higher quality. In Europe this process may become reality due to recent regulations agreed by the European Parliament for member states to provide migration statistics that comply with a harmonized definition. However, within Europe and other parts of the world an increase in population mobility, a reduction in administrative and regulatory barriers to movement and an increase in irregular migration have created greater pressures on the current ability for statistical systems to measure migration effectively. Inconsistencies are likely to occur for the foreseeable future and data collection methods may continue to struggle to capture movements.

In the context of the framework of this study, a greater amount of good quality data provided by national statistics institutes may improve both the estimation of correction factors for the harmonization of reliable flows and provide a more complete data set to estimate missing counts. Alongside better migration statistics, more documentation of summaries and comparisons of data, further improvement in the estimation of comparable migration flow data may be gained. The methodology presented in this thesis allows a great deal of flexibility for the estimates of comparable data from alternative regions and different size tables. Caution should be taken for flow tables of migration between nations that are very different, as models may struggle to explain all moves, especially those for political or legal factors. Additional dimensions for migrant characteristics such as age and sex might also be incorporated, if and when data become available. In addition, the methods developed in this thesis could be used for other measurement of transition between regions or states in which problems in inconsistencies and incompleteness occur.

This thesis has developed an estimation methodology for migration flow tables of comparable flow data between a set of countries. A concentration on two predominant factors, inconsistencies and incompleteness, were discussed and addressed via computationally intensive mathematical and statistical techniques. This allowed estimates of a complete table of comparable international migration flows that can be used by regional policy makers and social scientists alike to better understand population behaviour and change.

Bibliography

Abel, G. J. (2008, July). Modeling International Student Migrant Tables. *S3RI Methodology Working Papers M08*(05).

Agresti, A. (2002, July). *Categorical Data Analysis (Wiley Series in Probability and Statistics)*. New Jersey, USA: Wiley-Interscience.

Berndt, E. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement. 3*, 653–665.

Birch, M. (1963). Maximum Likelihood in Three-Way Tables. *Journal of the Royal Statistical Society, Series B 25*, 220–233.

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

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

Brierley, M. J., J. J. Forster, J. W. McDonald, and P. W. S. Smith (2008). Bayesian estimation of migration flows. In J. Raymer and F. Willekens (Eds.), *International Migration in Europe*, Chapter 7, pp. 149–174. Chichester, United Kingdom: Wiley.

Cameron, C. A. and P. K. Trivedi (1998, September). *Regression Analysis of Count Data (Econometric Society Monographs)*. Cambridge, United Kingdom: Cambridge University Press.

Carlin, J. (1987). *Seasonal Analysis of Economic Time Series*. Ph. D. thesis, Harvard University.

Castles, S. and M. J. Miller (2003, July). *The Age of Migration: International Population Movements in the Modern World* (3rd Revised edition ed.). London, United Kingdom: Palgrave Macmillan.

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

Cohen, J. E., M. Roig, D. C. Reuman, and C. Gogwilt (2008, October). International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences 105*(40), 15269–15274.

Congdon, P. (1991). General linear modelling: Migration in london and south east england. In J. Stillwell and P. Congdon (Eds.), *Migration Models: Macro and Micro Approaches.*, Chapter 7, pp. 113–136. London, England: Belhaven Press.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics.* London, United Kingdom: Chapman and Hall.

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

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

Deza, M.-M. and E. Deza (2006, October). *Dictionary of Distances.* Amsterdam, The Netherlands: Elsevier Science.

Erf, R. v. d. (2007, June). Feasibility study and associated work plan. Deliverable 1.2, Netherlands Interdisciplinary Demographic Institute (NIDI), The Hague, Netherlands.

Erf, R. v. d., L. Heering, and E. Spaan (2006a). Counrty report: Netherlands. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, pp. 553–564. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Erf, R. v. d., L. Heering, and E. Spaan (2006b). Statistics on asylum applications. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, Chapter 10, pp. 249–319. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Flowerdew, R. (1982). Fitting the Lognormal Gravity Model to Heteroscedastic Data. *Geographical Analysis 14*(3), 263–267.

Flowerdew, R. (1991). Poisson Regression Models Of Migration. In J. Stillwell and P. Congdon (Eds.), *Migration Models: Macro and Micro Approaches.*, Chapter 6, pp. 92–113. London, England: Belhaven Press.

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

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003, July). *Bayesian Data Analysis, Second Edition.* Boca Raton, USA: Chapman & Hall/CRC.

Giambattista, C. and M. Poulain (2006). Statistics on population with usual residence. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, Chapter 7, pp. 181–203. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

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

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

Guy, C. (1987). Recent Advances In Spatial Interaction Modelling: An Application To The Forecasting Of Shopping Travel. *Environment and Planning A 19*(2), 173–186.

Hardin, J. W. and J. M. Hilbe (2001, June). *Generalized Linear Models And Extensions.* College Station, USA: Stata Corp.

Herm, A. (2006a). Counrty report: Spain. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, pp. 633–643. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Herm, A. (2006b). Recomendations on international migration statistics and development of data collection at an individual level. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, Chapter 2, pp. 77–107. 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. (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., M. Jandl, and M. Hofmann (2006). The evolution of eu migration policy and implications for data collection. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, Chapter 1, pp. 35–75. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Kupiszewska, D. and B. Nowok (2008). Comparability of statistics on international migration flows in the european union. In F. W. J. Raymer (Ed.), *International Migration in Europe: Data, Models and Estimates.*, Chapter 3, pp. 41–73. London, England: Wiley.

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

Lee, E. (1966). A Theory of Migration. *Demography 3*(1), 47–57.

Little, R. J. A. and D. B. Rubin (2002, September). *Statistical Analysis with Missing Data, Second Edition* (2 ed.). Hoboken, USA: Wiley-Interscience.

Louis, T. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society, Series B 44*(2), 226–233.

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

Mayda, A. M. (2007, May). International migration: A panel data analysis of the determinants of bilateral flows. CReAM Discussion Paper Series 0707, Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London.

Mayer, T. and S. Zignago (2006). Notes on CEPIIs distances measures. *Centre d́Etudes Prospectives et d́Informations Internationales (CEPII), Paris.*

McCullagh, P. and J. A. Nelder (1989, August). *Generalized Linear Models, Second Edition.* London, United Kingdom: Chapman & Hall/CRC.

Mclachlan, G. J. and T. Krishnan (1996, November). *The EM Algorithm and Extensions* (1 ed.). Hoboken, USA: Wiley-Interscience.

Meilijson, I. (1989). A Fast Improvement to the EM Algorithm on its Own Terms. *Journal of the Royal Statistical Society B 51*(1), 127–138.

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

Nelder, J. and R. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Nocedal, J. and S. J. Wright (1999, August). *Numerical Optimization.* New York, USA: Springer.

Nowok, B., D. Kupiszewska, and M. Poulain (2006). Statistics on international migration flows. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, Chapter 8, pp. 203–233. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Orchard, T. and M. Woodbury (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 697–715.

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

Pedersen, P., M. Pytlikova, and N. Smith (2004). Selection or network effects? migration flows into 27 oecd countries, 1990-2000. Technical Report 1104, Institute for the Study of Labor, Bonn, Germany.

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

Perrin, N. and M. Poulain (2006a). Counrty report: Luxembourg. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, pp. 519–527. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Perrin, N. and M. Poulain (2006b). Counrty report: United kingdom. In M. Poulain, N. Perrin, and A. Singleton (Eds.), *Towards the Harmonisation of European Statistics on International Migration (THESIM)*, pp. 645–655. 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, Perugia, Italy. Joint Economic Commission for Europe (ECE) and Eurostat.

Poulain, M. and L. Dal (2007, March). Estimation of all flows within the intra-eu migration matrix. Deliverable, GéDAP-UCL, Louvain-La-Neuve, Belguim.

Poulain, M. and L. Dal (2008, June). Estimation of flows within the intra-eu migration matrix. Deliverable, GéDAP-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. In J. Raymer and F. Willekens (Eds.), *International Migration in Europe*, Chapter 10, pp. 209–234. Chichester, United Kingdom: Wiley.

Raymer, J. and G. Abel (2008, March). The mimosa model for estimating international migration flows in the european union. Joint UNECE/Eurostat Work Session on Migration Statistics, Unite. United Nations Statistical Commission And European Commission Economic Commission For Europe Statistical Office Of The European Communities (EUROSTAT).

Raymer, J., G. Abel, and P. Smith (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(4), 891–908.

Raymer, J., A. Bonaguidi, and A. Valentini (2006). Describing and projecting the age and spatial structures of interregional migration in Italy. *Population, Space and Place 12*(5), 371–388.

Rees, P. H. (1980). Multistate demographic accounts: measurement and estimation procedures. *Environment and Planning A 12*, 499–531.

Rogers, A. (1980). Introduction to Multistate Mathematical Demography. *Environment and Planning A 12*(5), 489–498.

Rogers, A. (1990). Requiem for the Net Migrant. *Geographical Analysis 22*.

Rogers, A., F. Willekens, J. Little, and J. Raymer (2002). Describing migration spatial structure. *Papers in Regional Science 81*(1), 29–48.

Salt, J. (1993). Migration and population change in europe. Technical Report 19UNIDIR/93/23, United Nations Institute for Disarmament Research, (UNIDIR), New York, New York.

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

Sen, A. K. and T. E. Smith (1995). *Gravity Models of Spatial Interaction Behavior (Advances in Spatial and Network Economics)*. New York, USA: Springer-Verlag.

Shimodaira, H. (1994). A New Criterion for Selecting Models from Partially Observed Data. *Lecture Notes In Statistics*, 21–21.

Skrondal, A. and S. Rabe-Hesketh (2004, May). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, USA: Chapman & Hall/CRC.

Stewart, J. (1941). An inverse distance variation for certain social influences. *Science 93*(2404), 89–90.

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

UN (1976). *Demographic Yearbook 1976.* UN.

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

Venables, W. N. and B. D. Ripley (2003, September). *Modern Applied Statistics with S.* New York, USA: Springer.

Wei, G. and M. Tanner (1990). Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika 77*(3), 649–652.

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.

Willekens, F. (1999). Modeling Approaches to the Indirect Estimation of Migration Flows: From Entropy to EM. *Mathematical Population Studies 7*(3), 239–78.

Willekens, F. and N. Baydar (1986). Forecasting place-to-place migration with generalized linear models. In R. Woods and P. Rees (Eds.), *Population structures and models. Developments in spatial demography*, Chapter 9, pp. 203–244. London, England: Allen and Unwin.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modelling.* London, United Kingdom: Pion.

Zeger, S., K. Liang, and P. Albert (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics 44*(4), 1049–1060.

Zipf, G. (1942). The Unity of Nature, Least Action, and Natural Social Science. *Sociometry 5*(1), 48–62.

Appendix A

S-Plus/R Code

A.1 Poulain Constrained Minimization

```
poulain <- function(M, nr, base)
{
    if(dim(M)[3] != 2)
        stop("M must be a array of dimensions n x n x 2")
    #tidy up data to exclude non-referee (nr) regions
    M[is.na(M)] <- 0
    x <- matrix(NA, dim(M)[1], 2)
    dimnames(x) <- list(dimnames(M)[[1]], c("r", "s"))
    M <- M[ - nr,  - nr,  ]
    n <- dim(M)[1]
    #create A
    A <- matrix(NA, c(n * 2 + 1), c(n * 2 + 1))
    A[c(n + 1):c(2 * n), 1:n] <- -2 * M[,  , 1] * M[,  , 2]
    A[c(n + 1):c(2 * n), c(n + 1):c(2 * n)] <- 2 * diag(rowSums(M[,  , 2]^2))
    A[c(n + 1):c(2 * n), 2 * n + 1] <- -0.5 * rowSums(M[,  , 2]) *
     (rowSums(M[,  , 1]) + rowSums(M[,  , 2]))
    A[1:n, 1:n] <- 2 * diag(colSums(M[,  , 1]^2))
    A[1:n, c(n + 1):c(2 * n)] <- -2 * t(M[,  , 1] * M[,  , 2])
    A[1:n, 2 * n + 1] <- -0.5 * colSums(M[,  , 1]) * (colSums(M[,  , 1]) +
     colSums(M[,  , 2]))
    A[2 * n + 1, 1:n] <- 0.5 * colSums(M[,  , 1])
    A[2 * n + 1, c(n + 1):c(2 * n)] <- 0.5 * rowSums(M[,  , 2])
    A[2 * n + 1, 2 * n + 1] <- 0
    #set up vector for constraints for corrections
    b <- c(rep(0, 2 * n), sum(apply(M, c(1, 2), max)))
    #calulate initial corrections
    xx <- solve(A, b)
    #correction factors by r and s
    x[ - nr, 1] <- xx[1:n]
    x[ - nr, 2] <- xx[c(n + 1):c(2 * n)]
    #normalisation setting a r value (base) to 1
    if(is.integer(base) == T) x <- x/x[base, 1] else x <- x
    y <- (matrix(x[is.na(x[, 1]) == F, 1], n, n, byrow = T) * M[,  , 1] +
     matrix(x[is.na(x[, 2]) == F, 2], n, n) * M[,  , 2])/2
    #average values for refereed countries
    n <- dim(M)[1]
    r <- matrix(x[ - nr, 1], n, n, byrow = T)
    s <- matrix(x[ - nr, 2], n, n)
    list(A = A, b = b, y = y, x = x, dist = sum((r * M[,  , 1] - s * M[,  , 2])^2/
     (M[,  , 1] + M[,  , 2]), na.rm = T))
}
```

```
poulain.comp <- function(M, nr, base)
{
    #obtain correction factors for refereed countries
    p <- poulain(M, nr, base)
    n <- dim(M)[1]
    x <- p$x
    r <- matrix(NA, n, n)
    s <- matrix(NA, n, n)
    r[,  - nr] <- rep(x[ - nr, 1], each = n)
    s[ - nr,  ] <- rep(x[ - nr, 2], times = n)
    #obtain correction factors for non-refereed countries
    x[nr, 1] <- apply(s[, nr] * M[, nr, 2], 2, sum, na.rm = T)/
     apply(M[ - nr, nr, 1], 2, sum, na.rm = T)
    x[nr, 2] <- apply(r[nr,  ] * M[nr,  , 1], 1, sum, na.rm = T)/
     apply(M[nr,  - nr, 2], 1, sum, na.rm = T)
    r[, nr] <- rep(x[nr, 1], each = n)
    s[nr,  ] <- rep(x[nr, 2], times = n)
    #averages of scaled data
    y <- (r * M[,  , 1] + s * M[,  , 2])/2
    list(y = y, x = x)
}

poulain.direct<-function(M, nr, base)
{
    #get original A from poulain function
    temp <- poulain(M, nr, base)
    #remove lagrange partial derivative
    A <- temp$A
    A <- A[ - dim(A)[1],  - dim(A)[2]]
    #replace with 0's and a constant (not too small)
    A[base - sum(nr < base),  ] <- 0
    A[base - sum(nr < base), base - sum(nr < base)] <- max(A)
    #obtain b
    b <- temp$b
    b <- b[ - length(b)]
    b[base - sum(nr < base)] <- max(A)
    #obtain x
    x <- temp$x
    n <- dim(M)[1] - length(nr)
    #calcualte r and s
    xx <- solve(A, b)
    x[ - nr, 1] <- xx[1:n]
    x[ - nr, 2] <- xx[c(n + 1):c(2 * n)]
    r <- matrix(x[ - nr, 1], n, n, byrow = T)
    s <- matrix(x[ - nr, 2], n, n)
    list(A = A, b = b, x = x,
     dist = sum((r * M[ - nr,  - nr, 1] - s * M[ - nr,  - nr, 2])^2/
     (M[ - nr,  - nr, 1] + M[ - nr,  - nr, 2]), na.rm = T))
}
```

A.2 Distance Functions for Constrained Optimization

```
ChiSq <- function(x, M1, M2)
{
    n <- length(x)
    a <- matrix(x[1:c(n/2)], dim(M1)[1], dim(M1)[2], byrow = T)
    b <- matrix(x[c(1 + n/2):n], dim(M2)[1], dim(M2)[2])
    sum(abs(a * M1 - b * M2)^2/(M1 + M2), na.rm = T)
}
```

```
Man<-function(x, M1, M2){
    n<-length(x)
    a<-matrix(x[1:c(n/2)], dim(M1)[1], dim(M1)[2], byrow=T)
    b<-matrix(x[c(1+n/2):n], dim(M2)[1], dim(M2)[2])
    sum(abs(a*M1-b*M2),na.rm=T)
}

Euc<-function(x, M1, M2){
    n<-length(x)
    a<-matrix(x[1:c(n/2)], dim(M1)[1], dim(M1)[2], byrow=T)
    b<-matrix(x[c(1+n/2):n], dim(M2)[1], dim(M2)[2])
    sqrt(sum(abs((a*M1-b*M2)^2),na.rm=T))
}

Can<-function(x, M1, M2){
    n<-length(x)
    a<-matrix(x[1:c(n/2)], dim(M1)[1], dim(M1)[2], byrow=T)
    b<-matrix(x[c(1+n/2):n], dim(M2)[1], dim(M2)[2])
    sum( abs(a*M1-b*M2)/(a*M1+b*M2) ,na.rm=T)
}

Cla<-function(x, M1, M2){
    n<-length(x)
    a<-matrix(x[1:c(n/2)], dim(M1)[1], dim(M1)[2], byrow=T)
    b<-matrix(x[c(1+n/2):n], dim(M2)[1], dim(M2)[2])
    sum( abs(a*M1-b*M2)^2/ (a*M1+b*M2)^2 ,na.rm=T)
}
```

A.3 EM Algorithm for Negative Binomial Regression Model

```
glm.nb.EM <- function(model, data, tol, max.it, z0)
{
    if(all(is.missing(pmatch(names(data),"y")))==T)
        stop("data must have a response column named y with some missing data")
    data$original <- data$y
    #Initial E-step with some unknown parameter set
    data$y[is.na(data$original)] <- z0
    z <- data$y[is.na(data$original)]
    #Initial M-step
    m <- glm.nb(formula(model), data, maxit = max.it)
    fit <- m$fit
    #Record convergence
    lik <- cbind(model$twologlik/2, m$twologlik/2)
    beta <- cbind(c(model$coef, model$theta), c(m$coef, m$theta))
    #Second E-step before loop
    data$y[is.na(data$original)] <- c(fit)[is.na(data$original)]
    i <- 2
    while(any(c(abs(beta[, i] - beta[, i - 1])) > tol, na.rm = T)) {
        m <- glm.nb(formula(model), data, maxit = max.it)
        fit <- m$fit
        data$y[is.na(data$original)] <- c(fit)[is.na(data$original)]
        z <- cbind(z, data$y[is.na(data$original)])
        lik <- cbind(lik, m$twologlik/2)
        beta <- cbind(beta, c(m$coef, m$theta))
        i <- i + 1
    }
    return(list(z = z, beta = beta, beta.se = beta.se,
     final.model = m, final.data = data, lik = lik, it = i))
}
```

A.4 Supplemented EM Algorithm

```
em <- function(beta0, model, data)
{
    #E step
    fit <- exp(model.matrix(model, data) %*% beta0)
    data$y[is.na(data$original)] <- c(fit)[is.na(data$original)]
    #M step
    m <- glm.nb(formula(model), data, maxit = 100)
    beta <- c(m$coef)
    list(beta=beta, m=m)
}

jac<-function(b.star, b.init, model, data)
{
    dm <- matrix(0, length(b.init), length(b.init))
    for(i in 1:length(b.init)) {
        #sequential replace each element of b.star with b.init
        b.temp <- b.star
        b.temp[i] <- b.init[i]
        #run one iteration of em with altered beta (a mix of b.star,
         with one element of b.init)
        u <- em(b.temp, model, data)
        #fill in the relevant dm row with rate of change
        dm[i,  ] <- c(u$beta - b.star)/(b.init[i] - b.star[i])
    }
    list(dm = dm)
}
```

```
sem<-function(b.star, b.init, model, data, tol)
{
    #get first and second dm for comparison
    b.star <- em(b.star, temp.mod, data)$beta
    dm <- jac(b.star, b.init, model, data)$dm
    storedm <- array(c(dm), c(dim(dm), 1))
    b <- em(b.init, temp.mod, data)$beta
    dm <- jac(b.star, b, model, data)$dm
    storedm <- array(c(storedm, dm), c(dim(dm), 2))
    #set up monitoring objects
    r <- 2
    err <- apply(abs(storedm[,  , r - 1] - storedm[,  , r]), 1, max)
    converge <- c(err > tol)
    print(converge)
    #estimate mapping differential depending on row (i) until all
     errors less than tolerance
    while(any(err) > tol) {
        for(i in 1:dim(dm)[1]) {
            #if given row is not converged estimate the mapping differential
            if(err[i] > tol) {
                #sequential replace each element of b.star with current b
                b.temp <- b.star
                b <- em(b, temp.mod, data)$beta
                b.temp[i] <- b[i]
                u <- em(b.temp, model, data)
                dm[i,  ] <- c(u$beta - b.star)/(b[i] - b.star[i])
            }
            if(err[i] < tol) {
                #if given row has converged, set row of dm to previous values
                dm[i,  ] <- storedm[i,  , r]
            }
        }
        r <- r + 1
        storedm <- array(c(storedm, dm), c(dim(dm), r))
        err <- apply(abs(storedm[,  , r - 1] - storedm[,  , r]), 1, max)
        converge <- rbind(converge, c(err > tol))
        print(converge[r - 1,  ])
    }
    return(list(dm = dm, dm.it = storedm, converge = t(converge)))
}
```