UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Social Sciences

Model-Based Estimates of UK Immigration

by

George Disney

Thesis for the degree of Doctor of Philosophy

February 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Demography and Social Statistics

Thesis for the degree of Doctor of Philosophy

Model Based Estimates of UK Immigration

George Disney

To fully understand the causes and consequences of international migration and its impact on characteristics of the population in the UK, researchers and policy makers need to build a reliable, useful and understandable evidence base. However, available information and statistics on international migration in the UK are uncertain and have limitations that firstly need to be fully understood. Inconsistencies in availability, definitions and quality mean that there is work to be done to better understand UK immigration.

As a result, there are two main aims in this thesis. The first is to use statistical models to make better use of all publicly available data and information in the estimation of UK immigration. The second is to understand better the amount and specific sources of uncertainty in the publicly available data. In order to fulfil these aims three statistical models have been developed, applied and the results have been interpreted.

The first, a log-linear model, represents the current state of the art with regard to UK immigration estimation. Three publicly available sources of data are combined in order to produce citizenship-specific estimates of UK immigration over time. Following this, a Bayesian approach is taken. Two separate Bayesian models are specified, which incorporate further auxiliary data and subjective judgements on data collection, in the model estimates.

In order to estimate the Bayesian models, useable prior judgments of the data are elicited. Analysis and interpretation of the Bayesian model-estimates of UK immigration, which include coherent expressions of uncertainty, provide the basis for substantive conclusions on the uncertain nature of UK immigration data.

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, George Disney

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research:

"Model-based Estimates of UK Immigration"

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. None of this work has been published before submission.

Signed: .....

Date: .....

Acknowledgements

This research was funded jointly by the Office for National Statistics and the Economic and Social Research Council. I acknowledge the help of the ONS in providing the International Passenger Survey data and feedback and support throughout the studentship.

I would also like to acknowledge the help, support and encouragement of my supervisor Dr. Jakub Bijak. I have been very lucky to draw on his experience and expertise, especially in Bayesian modelling of migration. Developing the models, eliciting the priors and making sense of the results, with his help and suggestions, was intellectually challenging and very rewarding.

Definitions and Abbreviations

A8 – The eight countries with low per capita incomes that joined the EU in 2004

CSO – Central Statistics Office (Ireland)

BGR – Brooks Gelman Rubin Statistic

DWP – Department for Work and Pensions

EEC – European Economic Community

EU – European Union

GLM – Generalised Linear Model

HE – Higher Education

HESA – Higher Education Statistics Agency

HoC PASC – House of Commons Public Administration Select Committee

IPS – International Passenger Survey

LFS – Labour Force Survey

LTIM – Long Term International Migration

MCMC – Markov Chain Monte Carlo

MSIP – Migration Statistics Improvement Programme

NHS – National Health Service

NI – National Insurance

NINo – National Insurance Number

ONS – Office for National Statistics

SMS – Special Migration Statistics

UK – United Kingdom

UKSA – UK Statistics Authority

Chapter 1: Introduction

1.1 Introduction

To fully understand the causes and consequences of international migration and its impact on characteristics of the population in the UK, researchers and policy makers need to build a reliable, useful and understandable evidence base. However, available information and statistics on international migration in the UK are uncertain and have limitations that firstly need to be fully understood. Inconsistencies in availability, definitions and quality (Poulain et al 2006) mean that there is work to be done to better understand international migration in the UK.

Researchers, academics, policy makers, the media and the general public are often interested in the levels of international migration to and from the UK; and often to and from specific countries. Also the public good of having a more detailed understanding of population dynamics cannot be understated. Having reliable statistics on the size and demographic characteristics of the population is vital for the provision of equitable and good public services.

The users and uses of population statistics are wide-ranging. They are used on a daily basis by policy makers, politicians, public health officials to help inform important decisions (Boyle and Dorling 2004) and by researchers to advance knowledge and to aid our understanding of society (Stone 1975). They can be used to simply to determine the population size within countries or regions, to establish the boundaries for political constituencies, for example. Or they are used in a more indirect way, providing denominators for other measures.

With such a demand and need for information on populations, trusted, independent and robust information about the size, structure and characteristics of a population are seen to be an essential underpinning of a modern society (Statistics New Zealand 2011). They are essential for improving the well-being, prosperity and legitimacy of modern democratic institutions and society alike. As such, it is vital that the statistics are not only reliable and robust, as previously mentioned, but also that users understand how the different statistics were compiled, how they relate to each other and what the uncertainty inherent in the estimates is. As such, it is vital to not only have reliable and trusted international migration statistics, but to also have a clear and documented account

of how these statistics are arrived at, their strengths and weaknesses and the concepts and definitions they are based on.

Society and its demographic structure are in constant flux, and over the last few decades have transformed in various ways, resulting in specific social challenges. In 2004, there was a significant expansion of the European Union, with the accession of 10 new member states, with two more following suit in 2007. Partly as a result of this, and the expansion of freedom of labour movement there has been a significant inflow of migrants to the UK from both within and outside the European Union, a key feature of demographic change and driver of the recent population growth (Boden and Rees 2010). There are also greater and greater levels of ethnic diversity in the UK in recent decades, with important regional variations in the ethnic composition and rate of change of the sizes of various ethnic groups (Rees and Butt 2004), which, to a certain extent, are driven by international migration.

The above are a few examples of the demographic and social context that motivates the study, collection and estimation of population statistics, all of which have important social policy implications and create research questions that need to be addressed. For population changes to be understood and social policy to be formulated in their response, it is important to have statistics on the size, location and characteristics of the population in the UK. Immigration to the UK and its estimation is an important component of this and is the focus of this thesis.

1.2 UK Immigration Statistics

Currently, the majority of UK international migration statistics are estimated by the Office for National Statistics (ONS) using the International Passenger Survey (IPS), a sample based survey used at ports and airports to produce estimates of immigration and emigration. International migration, in comparison to everyday international travel, is still a relatively rare event. This means that there are small sample size issues with the IPS and, as a result, there are irregularities in the data that are a reflection of sample noise and not a true reflection of international migration flows. These problems mean that the estimates of origin-specific migration flows are highly irregular over time (Raymer et al 2011 a). Furthermore, migration, in general, is the most volatile component of population change and is notoriously difficult to measure and define (Bijak 2010).

Consequently, this research is of great importance to the ONS Centre for Demography, who, as part of their recently completed Migration Statistics Improvement Programme (MSIP) (ONS 2012), are currently seeking new methods and sources of data to improve their existing international migration statistics. Moreover, as part of the MSIP, the ONS identified statistics on international migration as being in need of improvement.

The motivation for the MSIP also came from Government and Parliament who identified the need for more timely, accurate and detailed information on the UK population. For example, the Treasury Select Committee in its Eleventh Report (2008) made clear that international migration is the least reliable component of UK population estimates and the IPS alone is not an adequate source of data to estimate international migration in the detail and accuracy required by researchers and policy makers.

The MSIP is a cross government programme, which was set up to improve the understanding of migration statistics, sources and methods that underpin their production (Raymer et al 2012). Furthermore, the MSIP's remit to ensure estimates of migration – both internal and international – are more relevant to users' needs and as accurate as possible to ensure they are trusted as being the definitive and authoritative statistics on UK migration (ibid). The main outcomes of this research could make a substantial contribution towards improving our understanding of UK immigration statistics.

Further to the ONS other beneficiaries of the research include all potential users of international migration statistics in the UK. With the current statistics often being of a poor and unacceptable quality, the current research looks to combine available data sources that can help estimate international migration. In terms of methodology, this study research will also benefit researchers and statisticians who are interested in applied methods for improving population statistics, through the augmentation of alternative data sets in statistical models.

With the expansion of the European Union and the associated freedom of movement, there is an additional geographical and political dimension to international migration in recent years. A consequence of this is an increased demand for accurate, timely and comparable statistical migration flow data at the European level (EU 2007). Countries normally collect international migration data either according to their own needs or to be consistent with how they have collected data in the past (Raymer et al 2012). Until

recently there has not been international institutional pressure to collect international migration statistics that are internationally comparable.

The ONS now, though, have mandatory requirements to provide migration flows to Eurostat (the statistical office of the European Communities) following Article 3 of the European Parliament Regulation (EC) No. 862/2007. The following information is taken from Article 3 of the aforementioned regulation. Member states are required to supply the following international migration flow data to Eurostat:

a) Immigrants disaggregated by:

   (i) Groups of citizenship by age and sex;

   (ii) Groups of country of birth by age and sex;

   (iii) Groups of country of previous usual residence by age and sex;

b) Emigrants disaggregated by:

   (i) Groups of citizenship;

   (ii) Age;

   (iii) Sex;

   (iv) Groups of countries of next usual residence.

Alongside these requirements, member states of the European Union are also encouraged to supply more detailed origin/destination international migration data on a voluntary basis.

Following the regulation, the ONS is now obligated to provide reliable international migration statistics that are based on an internationally comparable definition; the UN definition of long term international migration. With inconsistencies in definitions and varying quality of data sources, it is recognised that, as stated in Article 9 of Regulation 862/2007, "As part of the statistics process, scientifically based and well documented statistical estimation methods may be used."

Consequently, in 2011 the Migration Statistics Unit in the ONS commissioned a report to deliver recommendations on how they could improve the estimates of international migration flows required by Eurostat (Raymer et al 2011 a). This report was limited in its scope to providing methodological suggestions to the ONS regarding how

they can fulfil the mandatory requirements of Eurostat. This thesis aims to go beyond the scope of the Raymer et al (2011) report, through the estimation of more detailed origin-specific international migration flows thus providing some of the more voluntary requirements of stipulated by Regulation 862/2007Eurostat, through the use of statistical modelling.

1.3 Thesis Aims and Scope

There are two main aims in this thesis. The first is to use statistical models to make better use of all publicly available data and information to produce estimates of UK immigration. The second aim is to understand better the extent and specific sources of uncertainty in the publicly available data.

In recent research on UK international migration statistics, there has been a focus on improving estimates of UK immigration and emigration (Raymer et al 2011 a). Raymer et al suggest that auxiliary administrative data can be combined with the main source of UK international migration data, the IPS, to smooth irregularities in the statistics and to improve estimates. Alternative sources of data on international migration are confined to immigration, with the only source of UK data on emigration being the IPS. As such this thesis focuses on the estimation of UK immigration between 2002 and 2010.

In the thesis, there is a slight shift of emphasis, in comparison to the most recent research on the estimation of UK immigration; which has specifically focussed on improving estimates. This aspect remains a focus of the current study, as making improved estimates contributes towards fulfilling the first aim of this research. However, moving beyond the previous research on UK immigration, understanding, considering and estimating the uncertainty inherent in UK immigration is an important contribution of this thesis. In the absence of collecting better data, it is necessary to use statistical models to not only to produce more certain estimates of migration, but to also better understand the uncertainty inherent in both the data and methods used.

As such, the main substantive contribution of the thesis is to outline three different approaches to using statistical models to estimate immigration into the UK and further our understanding of the uncertainty in the data. To do this, it is necessary to first gather together and assess all the publicly available information we have about immigration to the

UK. This is carried out in the first two substantive chapters of the thesis. In the subsequent four chapters, the development, results and implications of these three main modelling approaches are detailed.

1.4 Thesis Structure

The thesis is structured as follows. After this Introduction, Chapter 2 outlines the context of the research. There is a broad outline of the known long term patterns of immigration to the UK since 1950. Following this there is a review of the main theories of migration. Drawing on literature from different disciplines, there is a critical review of migration theory and its utility for estimation when there are issues with data reliability and consistency.

Chapter 3 presents a comprehensive audit and assessment of all publicly available data on UK immigration. The concept of true flow, based on the UN definition of a long term migrant is introduced. Following this a framework for assessing each source of data in relation to true flow, the quantity to be estimated, is outlined. The framework uses four data assessment criteria – definition, coverage, bias and accuracy – which are used to quantify how true flow is distorted by different characteristics of data collection.

In chapter 4 a log-linear model of immigration is detailed and results are presented. The main aim of the model is to combine sources of data at an aggregate level in a way which utilises their strengths and mitigates any weaknesses, identified in chapter 3. The model is based on the current state of the art in standard statistical modelling of migration. As such, there is a review of standard statistical applications in studies of migration where data is either unreliable or missing. This review motivates the specification and fitting of a log-linear model, with administrative sources of data included in the estimate via the use of an offset term.

Two Bayesian models of UK immigration are specified in chapter 5. Firstly, Bayesian modelling and its applications in related areas of research are introduced. From this review, and building on the conclusions of the log-linear model specified in chapter 4, a Bayesian Log-linear model is specified. The model specification allows for the judgements on further auxiliary data to be included as prior terms. Following this, the Data Assessment Model is specified. This modelling framework is designed to explicitly include, in the

model, the data assessment criteria detailed in chapter 3 and a coherent expression of the uncertainty inherent UK immigration statistics.

Having provided the framework which allows judgements of auxiliary data and the data assessment criteria to be included in a model estimate, chapter 6 discusses the process of elicitation of useable prior judgments of the data. These are based on concrete evidence, where available, and subjective judgement based on the audit and assessment of publicly available data in chapter 3. Furthermore, a method is developed which allows the uncertainty of each parameter to be calculated through the elicitation of a credible interval.

Chapter 7 details the main findings and implications of the results for the Bayesian Log-linear Model and the Data Assessment Model. Parameters of interest are analysed and interpreted, and estimates of citizenship-specific immigration are displayed. The uncertainty of each of the estimates is described and the sensitivity of each of the models to changing prior assumptions is assessed. Following this, appropriate methodological conclusions are made with regard to the implications of explicitly including subjective judgements of data characteristics, of varying levels certainty, in the estimation of UK immigration.

In the final chapter the main conclusions from the more substantive chapters are set out. Here, recommendations are made to ONS about data collection and possible approaches to improving our understanding of the uncertainty inherent in estimates of UK immigration. Furthermore, the key contributions of this research to the wider field of Demography, Social Statistics and the practice of international migration estimation are summarised and discussed.

Chapter 2 Research Context

2.1 Introduction

One of the main aims of the thesis is make use of all useful publicly available evidence to estimate flows of immigration into the UK. To produce model estimates of immigration to the UK, which attempt to overcome the inconsistencies and limitations of available data, it is important that all the relevant information about immigration to the UK is critically reviewed prior to any estimates being made. This chapter outlines what we know about the longer term patterns of immigration to and emigration from the UK and the theories that have been developed which help us understand what drives international migration.

With the absence of good quality data on UK international migration, it is important to have a sound understanding of the patterns of migration that are documented. This understanding helps to inform the assessment of the limited data that is available and is a prerequisite for any estimation later in the thesis. Furthermore, it will provide the basis for any verification of model estimates. Even though the focus of this thesis is on immigration, it is important to consider emigration from the UK to help understand how international migration flows between the UK and the Commonwealth, for example, have been established over time.

Even though there is a lack of detailed, good quality data on international migration flows, there has been a lot written about the main trends in UK international migration. Ranging from the Commonwealth migrations of the 1960s to the immigration following the expansion of the European Union in 2004, the general trends in migration are well-documented in the literature. Most of the evidence used to describe the main patterns of migration and what drives them is taken from census stocks of foreign nationals and the International Passenger Survey (IPS).

The first section of this chapter presents a review of the studies that use these sources of data to describe the changing patterns in UK international migration data after the Second World War. General patterns of international migration in and out of the UK after the Second World War are detailed. Providing the relevant historical context means that there is a good understanding of the development of international migration in the UK in recent years, so one can begin to establish what would be the expected general trends in

contemporary migration. In this section there is also a summary of the changing government policies and significant changes to international migration legislation, and how one would expect these changes to affect international migration to and from the UK.

To aid the understanding of what drives these general trends of international migration there is an established literature, which reviews theories that look to explain international migration (cf Massey et al 1993). However, there are established criticisms of international migration theory (cf Arango 2000).

The second section of this chapter reviews the relevant theory from the wider academic literature. The review of theory, alongside the main patterns of international identified, provides the foundation for the analysis of international migration statistics. Through this critical review, it is found that the utility of theory for this research is supplementary to the estimation process. International migration theory, because of its limitations, does not play a fundamental role in the estimation of later chapters. Rather, it is used to verify patterns and identify results that are unrealistic in the later chapters of analysis.

2.2 Known Patterns of UK International Migration 1950 - 2010

Since the 'age of mass migration', with approximately 55 million Europeans migrating to North America and Australasia between 1850 and 1914 (Hatton and Williamson 1998) scholars have striven to explain the phenomenon of human migration (Arango 2000). The UK has traditionally been a country of emigration; in the late 19th Century – 1870-1913 – the total net loss of the population due to emigration was 5.6 million, with the overwhelming majority of emigrants going to English speaking countries (Hatton and Price 1999). However, it should be noted that these figures include Ireland which was part of the UK at that time. With the American Immigration Acts of 1921 and 1924, restricting immigration, the numbers leaving the UK fell significantly.

More recently, over the last 50 years, the general pattern of international migration has changed. After the Second World War, with declining costs of travel, international migration revived (ibid). Traditional immigrant receiving countries – Australia, Canada and the United States – began to receive migrants from Asia, Africa and Latin America, rather than Europe (Massey et al 1993). This change is also evident in western and northern

Europe after 1945. Countries that were traditionally migrant-sending became migrant-receiving countries; initially from southern Europe, but later in the twentieth century from Africa, Asia and the Middle East (ibid).

This general pattern of global migration flows is evident in UK flows; traditionally a migrant sending country. The balance between immigration and emigration to and from the UK has changed since the 1950s. During the 1960s and 1970s emigration from the UK exceeded immigration to the UK (Hatton 2003). With the only comprehensive statistics on UK international migration coming from the International Passenger Survey (IPS) since 1964, Hatton and Price (1999) use the IPS to assess general trends in migration. They show that from 1964 until 1978[1] there were a greater number of emigrants leaving the UK and that 1979 was the first year that immigration was greater. However, as Hatton and Price point out, these figures do not include Irish migrants; estimates of Irish migratory flows have historically been obtained from the Central Statistics Office in Ireland. If one was to include the figures of Irish migratory flows then the balance between immigration and emigration would be altered. During the 1980s, the UK experienced net immigration and in the 1990s the level of net immigration increased. The following paragraphs outline what the main patterns were in the decades since the 1950s and, where the evidence is available, discuss the countries or regions where migrants both came from and went to.

During the period of net emigration outlined by Hatton and Price (1999) – 1960s and 1970s - the general destination of emigrants remained similar; mainly to the United States and to a greater extent to Australia, Canada, New Zealand and South Africa. However, care needs to be taken when assessing trends of net migration flows, used by Hatton and Price (1999). Net migration is a measure of the difference between immigration and emigration. As such, because immigration and emigration flows are not analysed separately, there is no way of knowing whether changes in net migration are as a result of a reduction or increase in either emigration or immigration (Rogers 1990). Furthermore, it could be the case that both emigration and immigration are increasing or decreasing, but at different rates.

From the 1950s, there were increasing levels of immigration to the UK from the Caribbean which was accompanied by increasing numbers from South Asia. The passing of the British Nationality Act in 1948, by the Atlee Government, meant that all "British

[1] Until 1963 the survey was called the Passenger Survey and excluded migrants travelling to and from European countries

subjects" had the right to enter the UK and enjoy all the social, economic and cultural rights of full British citizenship (Hansen 1999). Between 1948 and 1962[2], approximately 500,000 new Commonwealth immigrants had entered the UK, with the majority originating from India and Pakistan (ibid: page 95).

The immigrant flows from the Caribbean were started symbolically by the docking of the boat "Empire Windrush" in 1948. Caribbean immigration was driven by both recruitment pushes from the National Health Service, British Rail and London Transport and by family and island social networks that helped people establish their new lives in the UK (Byron 1994). This peaked in the 1960s and was effectively over, as one of the main flows of immigration to the UK, by 1973[3]. One possible reason for the slowing of immigration from the Caribbean could be that the population at origin of countries like Jamaica and Barbados, is relatively small in comparison to India, for example.

Flows of South Asian migrants to the UK have a long history; dating back to the seventeenth century (Peach 2006). In the years and decades after the Second World War South Asian migration increased dramatically. Previously, the main flow of migration from South Asia was of skilled Indian people; however, the increased flows from the 1950s onwards were numerically overtaken from people of a peasant background (ibid). Like the aforementioned flows from the Caribbean, these were mainly driven by demand for labour in the UK. The post-war boom in the economy meant that there was a demand for labour in the UK. This was not a new thing, every period of rapid economic growth previously had placed similar demands for new labour; previously taken up by migration from rural areas of the UK, Ireland and Eastern Europe respectively (Ballard 2002).

Pioneer migrants of the 1950s, who were taking advantage of available labour, went out of their way to assist fellow would-be migrants from villages back at origin (Ballard 2002). Surges of immigration were from India in the late 1950s, which peaked in the 1960s. During the 1960s there were increasing levels of immigration from Pakistan, peaking in the 1970s and then increasing levels from Bangladesh that peaked in the 1980s (Hatton and Price 1999: page 6). These respective specific waves of immigration to the UK, it is argued by Peach (2006), were a result of Partition of parts of South Asia from 1947 onwards. For example, the areas greatly affected by Partition – Indian and Pakistani Punjab and Pakistan-administered Kashmir – were major contributors to the emigrant flow to the UK. The

[2] 1962 saw the introduction of the *1962 Commonwealth Immigrants Act* which restricted immigration from new Commonwealth nations

[3] *The 1971 Immigration Act*, which further restricted immigration to the UK, came into force in January 1973

flows from India, Pakistan and Bangladesh, unlike the flows from the Caribbean, remain established, with persistent significant levels of South Asian immigration flows remaining (ibid)

The migration from South Asia in the 1950s and 1960s included mainly single men, labour migrants, who sent remittances back to their extended families (Peach 2006). However, following the 1962 Commonwealth Immigration Act, which attempted to restrict immigration to the UK, many of the male immigrants were faced with the issue of whether to bring their wives and families over to the UK with them. Sikh and Hindu men, according to Peach (2006), who were mainly from India, were the first to bring their families over to the UK. Followed by larger Pakistani family reunification in 1970s and Bangladesh family migration following 10 years later in the 1980s (Peach 2006, page 137). This pattern of firstly labour pioneer migration, followed by staggered family reunification from Indian, Pakistan and Bangladesh respectively, may go some way to explaining the aforementioned surges in immigration from South Asia. The family ties that this has brought, coupled with the large populations at origin of countries like Pakistan and India, mean that flows from South Asia are still significant today. For citizens outside the EU, close family members who already have leave to remain in the UK - such as spouses, parents and children – are entitled to apply to remain in the UK.

To understand the patterns of migration into and out of the UK it is also important to have an understanding of the wider European context of migratory flows. There have been major political and social transformations in Europe, since the 1960s, which have influenced patterns of migratory flows. Having an awareness of these will help aid understanding of European flows into and out of the UK, which are outlined by Jennissen (2004). Jennissen looks at each decade from the 1960s separately, firstly describing the migration trends qualitatively. Following this, he uses a multivariate analysis to help supplement qualitative descriptions of international migration trends with quantitative information. This analysis is based on an underlying expectation that there were a number of basic trends common to most European countries.

European international migration during the 1960s was mainly that of labour migration, according to Jennissen (2004), with a general south to north movement within Europe, with the exception of the Eastern bloc, where there was little emigration. The 1970s was a decade that saw change from the labour migration of the 1960s, with declining

emigration from Southern European countries, to increasing levels of family and return migration in Europe.

In this context of evolving European flows and following the aforementioned 'waves' of immigration from South Asia, Hatton and Price (1999) argue that immigration became more diverse from the 1980s onwards. This diversity of immigration can largely be attributed to increasing immigration from Europe, particularly from the European Economic Community (EEC) before 1993 and the European Union (EU) after the Maastricht treaty, with little evidence of increases in net immigration from elsewhere (Hatton 2003). However, this is also apparent across Europe too, with most countries becoming both significant senders and receivers of international migration (Castles and Miller 2009). Importantly, however, with regard to the continued use of the IPS as the main source of information on international migration flows, the UK and Ireland have opted out of the Schengen agreement. As a result, the UK has been able to maintain border controls, which enables the ONS to administer the IPS.

During the 1980s the average net immigration to the UK was 44,600 (Hatton and Tani 2005). The net immigration occurred mainly as a result of emigration falling; especially to Canada where net emigration fell from 148,000 in the 1970s to 47,000 in the 1980s, according to IPS estimates (Hatton 2005). Net immigration, however, continued in the 1980s from South Asian countries, with net immigration of 109,000 during the whole decade from Bangladesh, India and Sri Lanka combined. Concurrently with the aforementioned fall in net emigration to Canada, net emigration to Europe became net immigration.

In 1981 Greece joined the EEC and further expansion took place in 1983 with Spain and Portugal joining. This could have contributed to the increase in net immigration from the EEC. However, because the data used by Hatton (2005) is net flow of migration, there is no way of knowing whether the change in net flow was because of a reduction in emigration or an increase in immigration.

In the late 1980s there was an increase in numbers entering the UK and this continued in the 1990s. The level of net immigration to the UK continued to rise, increasing from 13,800 in 1993 - 1995 to 37,700 in 1996 – 1998 (Hatton 2005: page 724). This change in net migration was driven mainly by the total inflow of migrants following a rising trend, with an inflow estimated by the IPS in 1999 of 450,000 (Salt et al 2001).

Inflows of non-British people were also consistently higher than inflows of British nationals returning to the UK, according to the IPS, with the British inflows remaining relatively stable at around 100,000 per year. Non-British inflows, though, on the whole, rose in the 1990s. In 1990 the inflow of non-British migrants was 234,000; this rose steadily, apart from a slight dip in 1992 and 1993 to 331,000 in 1999 (Salt et al 2001: page 39). Within the non-British flow there was changing relative levels of immigration from different regions.

Salt et al (2001) use IPS data that is split into the following groups – EU countries, Old Commonwealth, New Commonwealth and Other Foreign countries. The main change in the late 1990s was a big increase in the in-flow from Other Foreign countries – from 76,500 in 1997 to 142,000 in 1999 (Salt et al 2001: page 42); whereas, migration from the EU, was estimated to have fallen in the late 1990s from 77,600 to 65,700 (ibid).

The dominant changes in the patterns of migration in the 2000s have mainly been driven by the expansion of the European Union. Following the signing of the Treaty of Accession in 2003 there have been three subsequent expansions of the European Union. The first came in 2004, the largest and most significant expansion of the two with regard to UK international migration. Ten countries joined the EU in May 2004 – Cyprus, Malta, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovenia and Slovakia. These countries, with the exception of Cyprus and Malta, are known as A8 countries, after their accession to the European Union. Following this large expansion a further two countries joined the EU in January 2007 – Romania and Bulgaria – and they are commonly referred to as A2 countries following their accession. Finally, Croatia joined the EU in 2013.

Post-enlargement, migration to the UK was found to be very different to migration of the past (Pollard et al 2008). One of the main findings of Pollard et al (ibid) was that it was both economically and logistically possible for people, who have migrated from the Eastern European accession countries, to come to the UK on a temporary or seasonal basis. In their paper Pollard et al (ibid) used the Labour Force Survey (LFS) to analyse both the quantity of migration from A8 countries and their economic activity in the UK. They found that in 2008 the LFS estimated that there were 665,000 A8 and A2 residents in the UK an increase of 550,000 from pre-expansion estimates. Furthermore, figures from the 2011 Census estimate that the number of usual residents who were born in Poland, in England and Wales, increased by 900% to 579,000 (ONS 2012 b)

The vast majority of this increase came from Poland who were according to Pollard et al, in 2008, estimated to be the largest foreign national group in the UK. Freedom of mobility to accession migrants can be seen in the opening up of transport links between the UK and Poland. Whilst in December 2003 40,000 passengers flew between only three British airports and Warsaw and Krakow in Poland by December 2007 one could fly from 22 British airports to ten different Polish cities (Pollard et al 2008: page 6).

Increasing migration from Eastern Europe during the 2000s was certainly different in character to the established flows from South Asia outlined previously. Furthermore, this represented an increase in more temporary migration, which probably was not as evident when the main flows were from further afield, such as South Asia and Commonwealth nations.

Another trend in recent years has been the increasing level of international students coming to the UK to mainly study at universities and Further Education colleges. Globally student migration flows are growing at a faster rate than overall international migration flows (King et al 2010). IPS estimates show that the total number of international students has been increasing since the early 1990s (Blinder 2014 a). India and China are the main countries of origin of international students (HESA 2010).

Although it is not a focus of this research, it is important to consider the inflow of asylum seekers to the UK. The number of asylum seekers in the UK rose from 4256 in 1987 to a peak of 84,130 in 2002 (Blinder 2014 b). Since 2002 these numbers have declined, a trend which is apparent across Europe, and in 2010 there were 17,916 (ibid). Whilst the total flow of asylum seekers to the UK is significant, it is below the European average. In recent years, the top asylum flows have included Iran, Iraq (peaking at 14,570 in 2002), Somalia and Syria (ibid).

In summary, the UK was mainly a country of emigration up until the 1980s, with 1979 being the first year of net immigration in the IPS data that Hatton and Price (1999) analyse. Patterns of emigration from the 1950s tend to mirror the traditional routes to English speaking countries of the Commonwealth. Immigration from the late 1950s is characterised by waves of migration from India, Pakistan and Bangladesh respectively and these particular flow have been less pronounced but still remain throughout the 1990s – 2000s.

From the 1980s onwards international migration has become more diverse and immigration from the European Union has become more important. The migrant flows from the Commonwealth, South Asia and the EU/EEC remained quite stable during the 1990s, with a large increase in immigration in the late 1990s followed by a further increase after EU enlargement in 2004. Over the last decade, it seems evident that the main fluctuations in immigration have been driven by changes in policy both at the national and EU scale which run alongside the more long-established flows from Commonwealth and South Asian countries.

2.3 Can Migration Theories be Used in Statistical Models to Estimate Migration?

The aim of this section is to review the main theories on the determinants and drivers of international migration and to assess whether they can be used in statistical modelling of migration. With the overall aim of the research focusing on using all available evidence to estimate UK immigration, through the use of statistical modelling, theory, if it is of direct use, needs to aid understanding of what affects levels and relevant characteristics of immigration to the UK.

Migration theory can be directly used in statistical modelling through the use of covariate information on characteristics that one believes could be driving migration. However, even if theory is not modelled explicitly, it can be used indirectly – making sense of complicated and sometimes contradictory patterns in data or to verify model estimates and rule out unrealistic results, for example. To determine the role of theory in the statistical modelling in this thesis, there is first an outline and overview of the main migration theories. This is followed by an account of the criticism migration theory has received in the literature.

2.3.1 Outline of Main Migration Theories

A fundamental challenge in the measurement and estimation of migration is that the aggregate spatial patterns of movement, represented in the data, reflect very complex combinations of motivations and reasons for international migration (Stillwell 2008). For example, the patterns are complicated by different levels of duration of stay (Boden and

Rees 2010). With this complexity, it is important to have an understanding of theories of migration to make sense of the flows represented in the data.

Ravenstein's (1885) paper "The Laws of Migration" tends to act as a starting point for work in migration theory that attempts to aid understanding of the quantity of migration (Lee 1966, Bijak 2010). However, apart from Ravenstein and Lee, there has been scant attempt to develop a coherent theory of the *quantity* of migration. Migration research tends to focus on the effect of migration on, for example, the economy (Coleman and Rawthorne 2004); on the characteristics of international migrants and whether this has a positive effect on skill level at origin or destination (Stark 2006), for example; or on the determinants of international migration (see Jennisen 2004). Whereas, using strands of theory from different disciplines to help aid improvements in the estimation of levels of migration has been somewhat neglected. This is part of the challenge that this section faces; using the literature on migration theory and the determinants of migration to aid the development of better estimates of international migration to the UK.

Following Massey et al (1993), firstly migration theories are considered that address the initiation of international migration. This is followed by a discussion of theories that consider its perpetuation once, theoretically, a migrant flow has been established. The initiation of migration flows can be conceptualised at various levels of analysis. From the individual level; as set out in the neoclassical theory of migration, where individuals look maximise their income (Borjas 1989), to the global level, as set out in dual labour market theory (Piore 1979) and world systems theory (Wallerstein 1974), where migration is part of a structural demand for labour or an inevitable consequence of a global economic system (after Massey et al 1993). Given that the data used in this thesis is aggregate in nature with relatively limited covariate information in the data sets, the review mainly looks at macro level theories.

The application of the neoclassical framework, which has been widely used in research contexts other than migration, marked a "watershed in the short history of thinking about migration" (Arango 2000:284). It is taken from the discipline of economics, and in this respect corresponds with the two sets of laws of migration set out by Ravenstein (1885) and Lee (1966) respectively; emphasising the primacy of economic drivers in migrations.

Neoclassical thought remains the dominant strand of economics and has played a considerable role in migration studies (Castles and Miller 2009). The neo-classical principle – the consideration of rational actors, cost-benefit calculations and the primacy of human agency – runs through many strands of migration research (ibid). As such, its impact on the migration theory literature should not be underestimated.

At a macro-level this theory was originally developed to explain migration of workers within the structuring context of economic development (Todaro 1969), where an economic theory of immigration analyses the allocation of labour across international boundaries (Borjas 1989). According to neo-classical theory, international migration is caused by geographical differences in the supply of and demand for labour (Massey et al 1993). This seems to make intuitive sense, the rational actor would move from a labour market where their chance of gaining employment is low to where chances are high. This could help explain the increasing number of migrants from countries such as Poland since the enlargement of the European Union in 2004, which allowed freedom of labour migration to the UK. Further to differences in availability of employment, it is argued that wage differential across space contribute to particular migration flows (Arango 2000).

The idea of the rational actor is also clearly present at a micro-level, but also includes personal considerations - such as the costs of transport, learning a new language and adapting to a new culture, for example (cf Stouffer 1960). Individual migrants evaluate the costs and benefits of migrating and make a rational decision about the destination of their migration based on which location will provide the highest benefit (Borjas 1989). As such human capital, such as language skills, education and knowledge are important drivers of migration at an individual level. The simple and compelling nature of this part of migration theory, offered by neo-classical macroeconomics, has "strongly shaped public thinking and has provided the intellectual basis for much immigration policy" (Massey et al 1993: 433).

As previously mentioned, the analysis in this thesis will not be conducted at an individual level. Therefore, it is debateable how relevant micro-level theories are to improving the estimation of migration at an aggregate level. For illustrative purposes, though, lets first consider the neo-classical theory of economics and how this has been applied in the study of migration. According to Borjas (1989), individual migrants evaluate the costs and benefits of migration and make a rational decision about the destination of

their migration based on which location will provide the highest benefit, in terms of their lifetime earnings.

Setting aside how one would operationalise this in a model, using a predominantly neo-classical theory to explain levels of international migration is insufficient as it does not explain "why so few people move, given the huge differences in income, wages and levels of welfare that exist among countries" (Arango 2000: 286). To a certain extent this criticism is fair; but, studies of international migration that operationalise neo-classical theory often consider various mitigating factors. For example, Borjas' (1989) work includes barriers to migration – such as costs of travel; financial barriers and immigration policies, whilst still being based on the neo-classical principle of rational choice.

Neo-classical interpretations of international migrations are very much based on an ahistorical, behavioural causal theory and fail to take into account structural causal theory. People are constrained and influenced by history (Portes and Borocz 1989). Even though there still seems to be individual economic reasons underpinning most migrations, a theoretical framework which is purely economic is not sufficient and does not reflect the lived reality of international migration (King 2002). People do not make decisions about migration that are completely divorced from culture, history and the development of societies (Castles and Miller 2009).

Empirical work has found ahistorical theory to be short sighted with regard to migration; it is rarely the very poorest that move to the very richest country (Hatton and Williamson 2002, Massey 1988). With history and culture being important to the development of migration flows, one could argue that awareness of past patterns of immigration to the UK takes on a much greater significance than a consideration of neo-classical theory of migration in a statistical model.

Furthermore, during the third quarter of the twentieth century there was rapid and sustained economic growth, globalisation of capital (Harvey 1999), decolonisation and economic development in the Third World. With this rapid economic growth, considerations of international migration tended to mirror the economic and political discourse of economic development, free markets and neo-liberalism; effectively a theory of migration could have developed as part of a specific, ideologically driven mode of economic growth. This is a longstanding concern, and was highlighted by Portes and Borocz (1989), who rightly make the case that migration theory might, as a result of its

formulation reflecting theory about the economic development of the time, lag behind reality. As such it is clear that other areas of migration theory need to be considered, to develop a more effective theoretical framework for aiding understanding of the quantities of UK international migration.

More recently than the long-standing application neo-classical economic theory in studies of migration, there has been the development of a strand of literature called the new economics theory of migration (cf. Massey et al 1993). The main similarity between neo-classical and new economic theories of migration are that they are both supply-side theories, in that they both focus on factors that impel people to cross borders in order to find work (Castles and Miller 2009). The key difference though is that decisions whether to migrate are not made by isolated individuals, but by larger groups of related people (Massey et al 1993). The groups' aims are to minimise risks and restraints associated with a variety of market failures. For example, a group may decide that one or more of their members should migrate, not just to get higher wages, but also to diversify income sources and to provide resources for investment in existing activities, such as the family farm (Castles and Miller 2009). This, to a certain extent, addresses the criticism that neo-classical theory is too individualistic, as the new economics of migration theory takes into account a migrant's position within and as part of a social group.

According to Stark and Taylor (1989, page 1) the potential gains in absolute income through migration are likely to play an important role in households' migration decisions. However, international migration by household members, which leads to success as labour migrants – such as higher earnings - can also be an effective strategy to improve a household's income position relative to others in the household's reference group. Whether this part of the relative deprivation hypothesis holds is debateable, as many migrants find themselves in a relatively low social class and menial work of the 'dual labour market' (Piore 1979); and, therefore, may not necessarily have improved their relative social position, albeit in a different society.

Empirical research has shown that, controlling for given income levels and expected income gains, a higher total relative deprivation of a population results in a higher levels of migration (Stark 2006) and households' initial relative deprivation will be directly and positively related to their propensity to send migrants (Stark and Taylor 1989). In this respect there is a link between the literature that looks at the effect of inequality as a

migration determinant and the new economics of migration literature outlined previously, as both consider migration as a decision that is not necessarily made at an individual level.

However, the relationship between inequality and migration is an example of the tension in migration theory between micro-level decisions and macro-level social structures as theoretical frameworks to explain levels of international migration flows. For example, in his studies of relative poverty Townsend (1985: 660) highlights the problems of trying to explain a social phenomenon using the minor theme of individual motivation rather than the major theme of social structures and organisation. Inequality is an intrinsically social concept, as it is a way of conceptualising and measuring particular relationships within a society. Townsend's (1985) insight can be applied to the study of inequality and migration; if one argues that there is any relationship between inequality and migration one is assuming that migration is socially structured in some way.

Up until now, most of theory outlined has mainly been supply-side theory that considers the characteristics of individuals. However, in the literature, there are theories which place primacy on the intrinsic demand for labour migration. One of the main examples of this is dual labour market theory (Piore 1979). The main difference of the dual labour market theory to neo-classical theory of international migration is that it addresses the "intrinsic labour demands of modern industrial society" (Massey et al 1993: 440) rather than individual level decisions. It is a labour demand-side consideration which takes into account the structuring nature of labour markets on international migration flows.

Piore (1979) argues that there is a permanent demand for immigrant labour that is inherent to the economic structure of developed countries (Massey et al 1993, Bijak 2010). A division into primary and secondary labour markets occurs (Castles and Miller 2009) where workers are selected on the basis of human capital and labour market segmentation. Viewing labour markets as being segmented allows you to conceptualise the allocation of jobs on ascribed characteristics (Peck 1996), as such it could be argued that there is the development of 'immigrant jobs' (Piore 1979) – certain occupations that are socially constructed as being carried out by an immigrant labour force. However, what constitutes an 'immigrant job' is place specific; in most continental European countries, jobs on assembly lines in car manufacturing came to be considered as 'immigrant jobs' whereas in the United States they are considered to be 'native jobs' (Massey et al 1993: 453).

This particular theory of migration can help aid our understanding of the patterns of migration, outlined previously, during the 1960s and 1970s. During this period there was increased immigration to the UK of people from Pakistan and Bangladesh, which was largely driven by the intrinsic demand for labour, particularly on night shifts, the least desirable work, in industrial towns of Lancashire and Yorkshire (Peach 2006).

With theories of labour market segmentation, the construction of a dual labour market and the social labelling of 'immigrant jobs', again it is clear that the determinants of migration are best not only understood using theory taken from the field of economics. International migration is socially as well as economically constructed. As such it is necessary to consider the contribution that theory taken from the study of social relations makes to our understanding of levels of international migration.

There is a long standing history of sociological considerations within the field of migration theory; but, there has been a lack of engagement with this theory. The literature, especially amongst demographers, is largely made up of empirical work, but this does not necessarily always aid easy-to-understand theories about the levels of international migration (Lee 1966). Similarities exist between certain migration theory taken from sociology and that of neo-classical rational choice, outlined previously. There are examples in both sets of theory where there is a cost/benefit model used to explain the determinants of migration; the classic sociological model of this nature being Stouffer's (1960) "intervening opportunities model". Stouffer's hypothesis is that the number of people travelling a given difference is proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities.

Where the two sets of theory do differ is the consideration here of more sociological determinants such as migrant networks (Massey 1990), world systems (Wallerstein 1974) and global cities (Sassen 1991); rather than purely labour market considerations - for example Borjas' (1989) consideration of the allocation of labour across international borders. These theories could be used to better understand the driving forces behind the South Asian migration of the 1950s and 1960s outlined in section 2.2.

The conditions that initiate international migration may be quite different from those that perpetuate it over time and space (Massey et al 1993). According to network theory, social migration networks increase the likelihood of migration (Massey 1990). Networks of migrants "interconnected by family or acquaintance ties assist subsequent

migrants in many aspects of everyday life" (Bijak 2010: 38). Taylor (1986), for example, makes the argument that networks of migrants at destination, who can identify with each other and have some kind of shared kinship, can act as a pull factor for international migration (cited in Massey et al 1993). This is because the networks act as a mechanism to lower the social costs of migration, as they provide assistance to migrants in everyday life. The theory of migration networks and their importance could be of great utility in helping to understand why there are specific migration flows into and out of the United Kingdom.

The idea of social capital and the 'strength of weak ties' has been developed within the wider sociological literature (cf. Putnam 1993) and it can be argued that the development of migration networks and institutions develop higher levels of social capital that drive migration flows. This can certainly be seen in recent immigration of Polish people to the UK and how networks are used to gain employment (cf Sumption 2009).

Migration network theory accepts migration as an individual or household decision, but also makes allowances for the importance of social relations, as well as economic considerations, in determining levels of international migration. It also allows for the impact of previous migration on contemporary and future migration flows. As such, network theory is not ahistorical; it places migration within an historical context. If one accepts the existence and importance of migration networks in determining the level of specific migration flows, it is clear that theory just taken from economics is insufficient. The networks may drive levels of migration that run against the assumptions outlined by neo-classical theory, for example.

There is an argument made that the unit of analysis should move away from being state-focused, towards cities and regions as the unit of analysis within a global economic system (Bloemraad et al 2008); world systems theory thus argues that international migration follows the political and economic organisation of an expanding global market (Massey et al 1993: 444). It makes the case that nation-states in less developed regions were incorporated into a global economic system that is controlled by 'western' capitalist nations (Wallerstein 1974). The controlling capitalist nations then look to exploit immigrant labour from the periphery for their continued accumulation of capital – global capitalism as a driving force behind international migration flows.

A specific example of the influence of global capitalism on international migration is Sassen's (1991) Global Cities thesis. She argues that the development of a bifurcated

service economy, characterised by increasing levels of immigrant labour, in cities such as New York and London, drives a certain type of migration flow. A mobile global elite occupies high-end service jobs in industry such as finance; whereas immigrants from less well-off countries move to take-up low-end service sector employment in occupations such as cleaning and working in restaurants. There has been much criticism of Sassen's thesis. It is argued that her theory is universally applied to global cities without being subjected to rigorous empirical work. Hamnett (1994) makes the case that Sassen universally applies the example of New York's reliance on immigrant labour to all 'global cities' without considering the specific nature of immigration to New York being different to that of London, for example. Furthermore, Sassen's study is focused on the specific cases of just a handful of cities, albeit large and influential cities, in relation to international migration. Interestingly, though, results from the 2011 Census do seem to show that London is a major receiver of international migrants with 24% of its population being made up of non-UK nationals (ONS 2012 b). Further research is required to see if Hamnett's criticism of Sassen's theory still holds.

Even though large amounts of migration theory are composed mainly of economic concepts, there have been attempts at using an international migration systems approach (Kritz and Zlotnik 1992) to synthesise a range of theory into one understandable framework. Migration systems theory originates in geography. A migration system is constituted by two or more countries that exchange migrants with each other. It suggests that "migratory movements generally arise from the existence of prior links between sending and receiving countries based on colonisation, political influence, trade, investment or cultural ties" (Castles and Miller 2009: 27).

The synthesis of theory has been a more recent development within the literature and is far from fully developed. It also accords central importance to global economic forces as drivers of international migration (cf. Jennison 2004, Massey 2003). It is argued that "international migration originates in the social, economic and political transformations that accompany the expansion of capitalist markets into pre-market and non-market societies" (Massey 2003, page 11). As a result, a theorisation of a migration system is place specific – the way that one migration system functions will be different to another. This is fundamentally different from the more universal nature of dual labour market theory; and definitely different from the ahistorical nature of neo-classical theory, as it explicitly takes into account the importance of past patterns of migration.

From this review of international migration theory, it is clear that an over reliance on purely economic explanations of migration is not enough to understand what drives immigration in general and more specifically to the UK. A consideration of a broad range of theories, and a historical perspective, is required if we are to create a theoretical framework which aids our understanding of the evidence available to estimate UK immigration. That said, in the literature, there are criticisms of the main theories outlined in this section. A discussion of these criticisms is the main focus of the following section.

2.3.2 Criticism of International Migration Theory

Firstly the utility of migration theory, per se, needs to be considered, in relation to the research topic. As outlined in chapter 1, there are large amounts of inherent uncertainty in the estimates of contemporary UK immigration, due to the nature of the data that is being used. Through the review of some migration theories above, it also clear that there are limitations and major obstacles to be overcome in the development of a theory that can aid the study of quantities of international migration. Therefore, it could actually be the case that introducing considerations of theory into the estimation of international migration, might only add further uncertainty. With this in mind, a more appropriate research strategy might be to approach considerations of theory slightly differently – leaning more towards using theory to both verify and rule out unrealistic patterns in the estimates provided by statistical models.

Secondly, there is no general theory of international migration and the theoretical base for understanding the forces shaping it remains weak (Massey et al 1993). This theoretical weakness exists within an academic context of there being considerable progress, in the second half of the Twentieth Century, in the understanding of the complexities of migration, which has often resulted from empirical research abstracted from theory (Arango 2000). We understand better the impact of migration on the economy in specific cases, remittances, integration into society, for example. The coherence of theory in relation to international migration measurement is therefore limited. This limitation, it is argued, is "part and parcel of the general difficulties that the social sciences experience when trying to explain human behaviour, affected by a large number of interrelated variables" (ibid: 295).

Another possible explanation could be that the large level of uncertainty and challenges there are in estimating international migration accurately and reliably mean that there has not been a good enough evidence base upon which to empirically test theories of international migration.

It is debateable, though, whether it is desirable – or indeed possible - to have a general theory of migration. It would be very difficult and perhaps unrealistic to try and develop a grand theory of migration, encompassing all of its drivers, characteristics and complexities. It has been argued that seeking such a synthesis of theory would be misguided; migration is too diverse and multifaceted a process to be explained by a single theory (Arango 2000). Rather, theories have developed largely in isolation from one another and are sometimes fragmented by disciplinary boundaries (Massey et al 1993).

In general, much of the main theories of the determinants of migration have predominantly been borrowed from the social science sub-disciplines of economics, sociology and geography (Massey et al 1993, Arango 2000). For theory to aid understanding of one's empirical research and for empirical research to aid theory building (thus attempting to address the criticism of Arango outlined above) it is important that one constructs a theoretical framework that is relevant to one's empirics.

As a result, it is necessary to select appropriate parts of existing theoretical work from the different disciplines within social science. This could then be used to make sense of the flows represented in the data, verify whether any estimates made potentially reflect the true level of international migration flows.

2.3.3 Discussion of the Role of Theory in Migration Estimation

From the review in the previous section, it is clear that there are strengths and limitations to each of the theories outlined. With no coherent theory of migration and with the fragmentation of migration theory across disciplines, this makes selection of appropriate covariate information problematic. For example, selecting variables which address the initiation and perpetuation of migration; that take into account migration systems and dual labour market systems at the same time as being culturally and historically grounded would result in an extremely complex statistical model, with very large – if not overwhelming – uncertainty associated with it.

It could also be argued that there is more certainty in what we know about data collection, definitions upon which migration data is based and coverage of varying sources of migration data, for example. Additionally, the information that we have about previous flows of migration in section 2.2 provides a better picture of the migration flows that we could expect to be evident in recent years, than the disjointed uncertain nature of migration theory.

Using past flows of migration as a guide to more contemporary flows is not completely abstracted from theory, though. Network theory helps us understand better the persistence of some of the more established flows from South Asia and the Commonwealth, for example. Furthermore, consideration of the documented past flows, addresses some of the criticisms of neo-classical theory of migration as we are viewing migration as something which is culturally and historically embedded.

Theory does help fill the gaps in empirical evidence, though, in statistical models and has been used to estimate migration where data is missing (Raymer et al 2013). Furthermore, Cohen (et al 2010) advocate extending previous research on gravity models (Kim and Cohen 2010). However, it is argued by Bijak (2010), that prediction based directly on theories of migration is not an option due to their fragmentation and associated uncertainty.

As such, the statistical models used in the thesis to estimate immigration to the UK, look to include considerations of what we know about the data, its limitations and strengths and differences in characteristics of data collection of the different sources available. They do not explicitly include covariate information selected based on migration theory. Past flows, outlined in section 2.2 will help verification of estimates and migration theory will allow us to make sense of the estimates and to rule out unrealistic results.

2.4 Conclusion

This chapter has outlined the main patterns in migration and the main theories in the international migration literature. From section 2.2, it is clear that immigration to the UK has two main features. Firstly, there are the long established flows to the UK – from countries such as Australia, Canada and New Zealand which seem to provide a significant source migrant inflow. Furthermore, since the 1950s, there has been significant

immigration to the UK from South Asian countries such as India and Pakistan, which seems to persist. Secondly, there is the increased level of immigration from Eastern Europe, especially Poland, as a result of the enlargement of the European Union. On top of these established flows, increasing student inflows have led to China being a significant sender of immigrants to the UK. Furthermore, and whilst estimates of asylum and refugee flows are not a focus of this thesis, countries which have been affected by conflict and unrest, are also significant senders of asylum seekers and refugees to the UK.

With regard to international migration theory, because of its limitations and fragmented nature outlined in section 2.3, it is not clear that its use, in a model, will reduce the large amount of uncertainty associated with international migration statistics. As a result, of these limitations of international migration theory, the research in this thesis takes the form of a mainly data driven approach, where theory is used to both verify estimates and to rule out results which seem unrealistic. The statistical models are limited to combining various sources of data, and what we know about data collection, in the estimation of UK immigration. Consequently, the next stage in the thesis, in chapter 3, is an audit and assessment of all the available sources of data that can be used to estimate UK immigration in various statistical models.

Chapter 3 – Data Audit and Assessment

3.1 Introduction

This chapter outlines and assesses the main sources of data used in the thesis to estimate UK international migration. As outlined in chapter 1 measuring migration is a particularly difficult task. Chapter 2 established that a data-driven approach is going to be used to estimate international migration, as the application of theory directly in the estimation process is problematic. Consequently, from this chapter onwards, the main focus is the assessment of the available data, with a view to including these considerations, alongside the data itself, in the model-based estimates of immigration to the UK.

Many sources of data are not designed to capture migration specifically; rather they are used for other administrative purposes and contain information that can be of some potential use to estimate migration. This has implications for the quality of the data in relation to what users want regarding migration measurement. Furthermore, migration data is often unreliable or incomplete (cf. de Beer et al 2010). As such, when considering appropriate data for this thesis, there are inevitably a number of obstacles to overcome (Raymer et al 2012).

The main purpose of this chapter is to understand fully the sources of data that are used throughout the thesis and to identify the challenges that each source poses with regard to making the best use of the available data in estimation. To do this the concept of 'true flow' is introduced, with each of the data sets available assessed in relation to this. 'True flow' is based on the UN definition of long-term international migration. It is the quantity of migration flows by the variables of interest, if one had a perfect system to measure UK immigration. As a result of a discussion of the United Nations definition of an international migration, and a consideration of the sources of data available, four data assessment criteria are proposed.

Following this there is a review and high-level assessment of the available sources of UK migration data; starting with data the ONS predominantly use, mainly the International Passenger Survey, and then moving on to all other alternative sources of publicly available data. Data collection, the original purpose of the source and the definitions and concepts that they are based on are outlined. Assessment of each of the

sources in relation to true flow is conducted using the four assessment criteria outlined earlier in the chapter.

3.2 "True Flow" of International Migration

In existing research on estimating international migration, there is often a reference to 'true flow' (cf. Wisniowski et al 2011, Raymer et al 2011 b). Here true flow is the unknown number that is being estimated and represents the number that one would obtain if one was able to monitor a given definition – i.e. the UN definition (see 3.2.1) - of immigration and emigration perfectly, without bias and undercount and with complete coverage of the population. There is no data collection system which provides perfect information on true flow, as a continuous observation of all members of a population is not possible. As a result, in practice, all observations of migration are incomplete (Willekens 1999). Firstly, one must outline how true flow is defined throughout this research.

3.2.1 UN Definition

Migration is hard to define, thus making it a very difficult concept to measure. Numerous typologies of migrants have been produced, mainly based on distance moved, time spent away from origin and motivation for the migration (Salt et al 2001). Further to this, motivations for migration are not immutable (ibid). Migrants might move in and out of the labour market and migrants who move to join up with other family members often end up working. The consideration of the fuzzy boundaries between migrant types is important for subsequent sections of this chapter, which consider variables included in the International Passenger Survey in section 3.3.1 and compare alternative sources of data in section 3.4.

Throughout her work on international migration data Zlotnik (1987) advocates homogeneity in the concepts underlying flow statistics on international migration. The homogenous concept of international migration that Zlotnik discusses has resulted in the UN definition of an international migrant, which is the working definition of true flow throughout this research. The United Nations Definition of a long-term migrant is as follows:

> *"A person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes his or her new country of usual residence. From the perspective of the country of departure the person will be a long-term emigrant and from that of the country of arrival the person will be a long-term immigrant" (UN 1998: page 18).*

To be able to assess data in relation to true flow, it is vital that one has a sound understanding of the UN definition of long-term international migration. Through a good understanding of this definition the strengths and weaknesses of the various sources of data used in the thesis become apparent.

The UN definition has developed and changed over time; in 1976 the Statistical Commission of the United Nations adopted "Recommendations on Statistics of International Migration" (Kraly and Gnanasekaran 1987, after UN 1980). The 1976 recommendations, because of their avoidance of the term "usual residence" and reliance purely on duration of stay at origin or destination, caused confusion amongst users and were not implemented consistently (UN 1998). Consequently, the UN definition was simplified. "An international migrant is defined as any person who changes his or her country of usual residence" (ibid: page 17). 'Usual residence' refers to the place that the person usually lives; where he or she normally spends the daily period of rest. Travel abroad – on business; visiting friends or relatives; going on holiday, for example – do not involve changing one's country of usual residence and thus do not count as an international migration.

In reality, international migration is a complex process (King 2002). For example, someone might migrate to another country for several months a year for work and return to origin after a period of time has elapsed. Seasonal workers may spend three months a year working abroad, and nine months at their place of usual residence. These examples would not constitute a long term migration in accordance with the UN definition. These types of circular migration and regular change of address, however, affect both the way people view their usual place of residence and duration of stay in a country. Within the UN definition, there are both spatial considerations - crossing of an international border - and a time consideration - to distinguish between an international visitor and an international migrant.

3.2.2 Data Collection as a Distortion of True Flow

The reliability and quality of international migration statistics are conditioned by the data collection systems used to capture meaningful changes in the country of usual

residence (Raymer et al 2012). The main sources of data of international migration are sample surveys; population registration systems both centralised and local; administrative registers related to foreign nationals, alien registers, residence permits, visas information and asylum seeker data bases; statistical forms filled in for all changes of residence; and census data (ibid: page 45). All of these sources of information measure slightly different things and in some cases are not intended to measure international migration. This section makes the case that each data collection system, in the context of improving international migration statistics, should be viewed as a distortion of true flow outlined above. Understanding the nature and the uncertainty of this distortion is the essence of the substantive analysis in this thesis.

In general migration data collection systems give rise to two main types of migration data – transitions and events. The difference between transition and event data is similar to the distinction between migrant data and migration data (cf. Willekens 1999). Transition data is generated when the data collection system requires information on residency at two separate points in time; if there is a change in the place of residence, at the two points in time, then a migration will be recorded; transition data is commonly captured in censuses. Event data is collected when the actual migration is recorded. For example, if one has to register a change of address, in a new place, this would constitute event migration data.

Not only is the type of data generated dependent on the collection system, there are also differences in the definition of international migration between countries. This makes direct comparisons of reported international migration data between countries impossible. Another common limitation to migration flow data, which often make international migration statistics unreliable and inadequate for use, is the under- registration of migrations, where the data collection system relies on self-declaration of movement. This is not necessarily a major problem in countries that have compulsory population registers. However, where administrative data is being used in the place of a register, this is a significant limitation for migration data users. Here, the distortion of true flow coming from the under-registration of people to a given administrative data source, is an undercount of migration. Furthermore, certain groups – young males, for example – may be more susceptible to this kind of distortion.

Data coverage, where the data collection system may not completely cover the correct target population and might systematically exclude some subsets of migrants, is a

further distortion of true flow. For example, students or asylum seekers might only be captured or excluded by the data collection process (Raymer et al 2012). The result of this is a systematic undercount of a certain group of migrants. Consequently, users of international migration statistics need to be fully aware that data from different countries, and from difference sources within countries, need to be treated with care.

Reliability of migration data can be defined as how well the measurement, or data collection, corresponds to the desired definitions and concepts (De Beer et al 2010). With differing definitions, though, reliability does not necessarily lead to comparability between data sets. As such, it is distinctly possible, that one can have two sources of information that are reliable, yet because of their conceptual basis, are not comparable. This nuanced consideration, of the extent to which data is suitable for measuring international migration in accordance with the UN definition, leads to the following assessment criteria of international migration data, in relation to true flow.

3.2.3 Data Assessment Criteria

Each data set that is available and contains information on international migration has advantages and disadvantages. A central consideration of this research is gaining an understanding of in what way and by how much true flow of immigration is distorted by the data collection process of each of the available sources. To carry out a systematic audit and assessment of the available data four data assessment criteria are outlined below. These criteria provide a framework that can be used to evaluate the distortion of the available data in relation to true flow. They are:

(i) **Definition** – how closely does the data match the UN definition of international migration?
(ii) **Coverage** – theoretically what portion of the total immigration flow does the data set cover?
(iii) **Bias** – is there any systematic bias as a result of the way the data is collected?
(iv) **Accuracy** – with regard to its intended purpose, how accurate is the data?

There is not a single data source that exists which matches the UN definition of international migration, is an accurate and comprehensive account of the all the different 'types' of migrations and where there is no systematic bias in how the data is collected.

However, there are, in respect to the assessment criteria outlined above, strengths and weaknesses in each of the data sets available. For example, survey data specifically designed to estimate migration in a given country will match the definition criteria of true flow well. However, with regard to accuracy, the sample survey would produce estimates which are subject to sampling variability. The issues with sample size mean there are often large standard errors around estimates; with only very large flows having acceptable levels of uncertainty. Additionally, there could be systematic bias in the data from respondents as immigration is such an uncertain and personally sensitive phenomenon.

It is highly possible, whilst using administrative data that data sets, even though they are very accurate, do not closely align with true flow. Administrative data, as it is not necessarily designed to collect information on migrants, may only cover a sub-section of migrants. So, with regard to coverage, administrative data sources might not a good reflection of the true flow of migration.

To aid the assessment of data, and thus contribute towards the understanding of how we can improve data reliability, it is important that one considers and outlines the meta-data of the data-sets being used (Poulain et al 2006). Where meta-data is not available or, in the case of administrative sources that are in general collected for a purpose other than estimating migration, it is important to consider carefully where the data comes from and the practical administrative procedures involved in the creating the data set (ibid). The following section considers this, in relation to the assessment criteria outlined above, for the data that is available for the estimation of UK immigration.

3.3 ONS Estimate of Long Term International Migration

There is no single, comprehensive system that the ONS uses to estimate UK immigration, as such the ONS uses a combination of data from different sources (ONS 2011 a) to obtain an estimate that has sufficient coverage. Each source of data contains different information that the ONS use to estimate UK immigration. The 'Long-Term International Migration' estimate (LTIM) is currently the most comprehensive estimate of long-term international migration in or out of the UK (ONS 2011 b). It is important to note, though, that none of the data used to estimate LTIM are specifically designed to solely measure international migration.

There are a number of different components that are used by the ONS to estimate international migration. LTIM is mainly based on the migration sub-set of the International Passenger Survey (IPS), which is discussed in more detail in the next subsection. It is important to note, however, that the only component of LTIM that is available to this research is the International Passenger Survey. Each of the components listed below briefly detail the current method of estimating LTIM; further details about these and the main changes to them over time are detailed below:

- International Passenger Survey (see section 3.3.1)
- Non-International Passenger Survey Components (see section 3.3.2):
    - Home Office administrative data on asylum seekers and their dependents who are not counted by the IPS;
    - An adjustment for 'visitor switchers' - people who intend to enter or leave the UK for less than 12 months but will actually stay or stay away for longer, this is obtained through questions in the IPS;
    - An adjustment for 'migrant switchers' - people who intend to enter or leave the UK for at least 12 months without those intentions being realised, this is also obtained through questions in the IPS;
    - Quarterly Irish National Household Survey data on flows to and from the Republic of Ireland provided by the Irish Central Statistics Office (used in 1991-2007 LTIM estimates, from 2007 flows from the Republic of Ireland have been estimated using the IPS);
    - Northern Ireland Statistics and Research Agency data on migration to and from Northern Ireland (used from 2008 onwards for LTIM estimates, before 2008 the IPS was used to estimate migration to and from Northern Ireland) (ONS 2011 b)

3.3.1 International Passenger Survey

The main, official source of numbers on immigration to the UK comes from the International Passenger Survey (IPS). The IPS is a sample survey of passengers arriving at, and leaving air and sea ports and the Channel Tunnel (ONS 2011 a). Originally the IPS when it was established in 1961 was designed as a passenger survey, gathering data on the impact of travel expenditure on the UK economy, impact of international tourism on the

UK and how this has all changed over time. There has since been questions added to the survey about international migration. Travellers are selected for interview and all interviews are conducted on a voluntary and anonymous basis. In 2010, 316,000 interviews were recorded and the response rate was 81% (ONS 2010). Even though this seems like a large sample, it is worth bearing in mind, that this amounts to approximately only 0.2% of a total of approximately 158 million travellers and of this sample only 2990 were migrant interviews (ONS 2010).

The migration sub-set of the IPS forms the main part of the LTIM estimate each year. The flows estimated from the IPS are gathered from face to face interviews that take place at ports around the UK. The IPS uses a multi – stage sample design, where specific sea crossings or times shifts at airports are selected and then travellers are chosen systematically at fixed intervals (ONS 2011 a: page 21). At airports a certain number of shifts are sampled randomly each quarter, stratified by time of day and day of the week; passengers are counted as they pass a predetermined line and every nth passenger is interviewed (ibid).

Sampling is similar at sea ports, where passengers may be sampled on the quayside as they embark or disembark with the timing of the interview shift selected at random (ibid). At some sea ports where, due to the layout of the port, sampling is not possible on land, some of the sampling is carried out on the boats and passengers are sampled systematically en route (ONS 2010). Finally, tunnel routed foot passengers are sampled in a similar way to airports, whereas people in vehicles crossings are randomly selected and the interviews take place on board the train itself (ibid).

The data taken from interviews need to be weighted to provide estimates of international migration. The calculation of these weights is extremely complex and needs take into account the information outlined above regarding the complex sample design. Total passenger traffic is provided by the Civil Aviation Authority, Department for Transport, Eurostar, Eurotunnel, BAA and a number of the airports.

The remaining part of this section is given over to assessing the IPS data in relation to true flow using the four data assessment criteria outlined in section 3.2.3.

Definition

It is worth considering, why the ONS continue to use the IPS to estimate migration flows. One of its strengths is that the questions regarding international migrants are designed with the UN definition of long term international migration in mind (Boden and Rees 2010). It consequently captures immigration flow data, which corresponds to the component of the UN definition about intention to migrate for 12 months. With regard to the definition assessment criteria, the IPS is a relatively close match to true flow.

Coverage

Before 2009 the main UK airports – Heathrow, Gatwick and Manchester – were always included in the sample. After 2009, though, changes were made to the sample design and data processing of the IPS (ONS 2011 a). The changes consisted of increasing the number of shifts at ports and airports except Heathrow where the number of shifts was decreased. New shifts were established at Aberdeen and Belfast airports with the Portsmouth to Bilbao sea port also getting new shifts. In general the smaller ports were included in the sample prior to 2009 if they had over 1 million passengers travelling through them each year.

Also, prior to 2009, extra samples were made – migrant filter shifts – on the inward flows at the four Heathrow and two Gatwick terminals to boost the sample of migrants and were first run in outflows in 2007. Passengers, who were contacted in these migrant filter shifts, were asked questions to simply identify if they were migrants or not and only migrants were given a full interview. This system of migrant filter shifts were abolished after 2009, with the sampling interval altered as follows – primary sampling interval for screening migrants, normally around 1:10 with a further interval of around 1:30 looking to carry out the full passenger interviews (ibid: page 22).

The aims of the changes outlined here are to ensure that the IPS becomes more 'migrant focused' and balanced in terms of the routed that migrants use; with a slight shift of focus moved away from Heathrow (ibid). As a result, currently for air routes, the 12 busiest sites – 5 terminals at Heathrow, 2 terminals at Gatwick, 3 terminals at Manchester International Airport, Stansted and Luton – are sampled regularly over the year (ONS 2010). A small number of shifts every quarter are also conducted at some of the less busy international airports. With regard to sea routes, ports with carrying over 50,000 passengers a year are normally included in the IPS sample (ibid). The ONS state that this means that

there is 95% coverage of passengers entering and leaving the UK by the IPS, with the missing 5% attributable to night time travellers and routes that are too small in volume to be deemed cost-effective to cover (ibid). Consequently, it is clear, that with regard to the coverage assessment criteria, as the sampling attempts to cover ports across the whole of the UK, the IPS is relatively good.

However, recently, there has been concern that, following the expansion of freedom of movement in the EU, that regional airports remain under sampled. Migrants from Poland, for example, are more likely to travel through smaller ports and airports, rather than the main hubs ONS (2012 c). This lack of coverage could lead to an undercount of immigration.

Bias

With regard this data assessment criteria, the main focus is on how interviewees could potentially respond to the survey in a biased way. Data collection and how data are collected are key to their reliability and utility. Accordingly the ONS place great emphasis on the training of their interviewers for the IPS (ONS 2010). As previously mentioned, the interviews are undertaken on a face-to-face basis and are firstly recorded on paper; they are later transferred to a computer system. Data collection and the associated practical problems are an issue for the IPS. For example, it is ONS policy not to interrupt telephone calls of passengers, so even though they are included, if selected, in the IPS sample they may not be interviewed (ONS 2010). This has the potential to introduce a small amount of bias; however, it is hoped that this reduces the chances of biased responses due to the inconvenience of interrupting someone's telephone conversation.

The IPS is an intentions-based survey, which means people are asked about whether they intend to migrate with the intention of making the UK their usual place of residence for a period of 12 months or more. With the IPS estimates being based on intended rather than actual duration of stay, the data might be subject to bias (Bijak 2010). For example, there might be respondent groups, from specific countries of origin, for instance, who at the point of interview are not sure whether it is their intention to immigrate to the UK due to the nature of work available to them, but who then go on to become long term international migrants.

Accuracy

All surveys are subject to random sampling variability; if many samples were made, estimates of the characteristics of migrants would vary. Accuracy of the IPS, with regard to sampling variability, will be assessed using a strategy similar to that of Raymer et al (2011). One would expect the larger flows – total immigration, totals cross tabulated by age and sex – to be relatively stable over time. Where there are erratic changes in the flow of immigration over time this is more likely to be a reflection of sample noise. There is an established literature on the strong regularities of migration by age patterns (cf. Rogers and Castro 1981). As such, through looking at the age schedules of IPS data one can obtain an indication of sampling variability; this is an approach used by Raymer et al (2011) and it is applied briefly here too.

In general, the smaller the number of interviews a variable is based on, the greater the level of variability in the estimate. Bearing in mind the aforementioned information on the sample of design and varying frequency of interviewing shifts across the country, this is another source of sample error. As such, because of small sample size issues and the sample design, the ONS only deems certain elements of IPS estimates as being accurate. These tend to be the larger migrant flows, as the estimates are based on a larger proportion of passengers sampled (ONS 2011 a). In relation to the assessment criteria, it is clear that true flow will be distorted in relation to accuracy. Consequently, one would expect smaller flows to exhibit large amounts of sampling variability with larger flows being more stable over time.

To illustrate this sampling variability a brief exploratory analysis is carried out. Figure 3.1 shows the estimate of immigration of foreign nationals to the UK for 2002 – 2010; there is a general upward trend and the pattern in the data seems relatively stable over time.

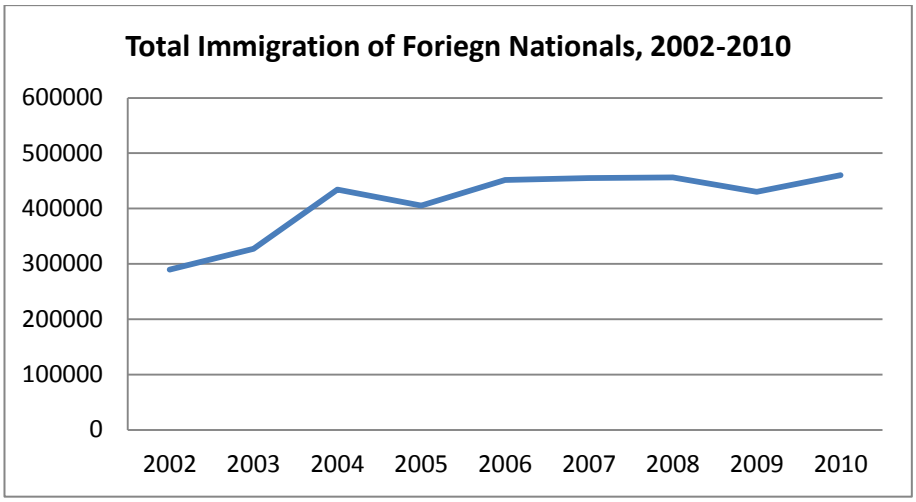

Figure 3.1: Total Immigration to the UK of Foreign Nationals, 2002-2010
Source: IPS

Immigration of foreign nationals are based on a large part of the sample of IPS interviews – between approximately 85% and 90% of all immigration interview contacts - and can therefore be treated as relatively reliable if one assumes that the IPS has good coverage of the inflow of migrants to the UK.

The age schedules of foreign immigrants in figures 3.2 and 3.3 illustrate a very regular pattern over time, which is similar to what one would expect from the literature on age schedules (cf Rogers and Castro 1981). There are zeros in the older age categories, which one might expect on occasion for the 90+ age group; but it is unlikely that there were not any immigrants aged between 70 and 74 in 2002 – 2004. The zeros could be explained by sample size issues; however, there is no way of confirming this is the case. Each schedule, except for 2002, peaks during the 20 – 24 age group.

Seeing as though, the main aim of the research is to develop methods to improve the reliability of international migration estimates, after the diagnostic test of looking at age schedules, the next port of call is to assess any calculated estimates of uncertainty. This should aid understanding of whether the patterns in the data are reflections of uncertainty and sampling error or actually reflect the true patterns of immigration into the UK. To determine the reliability of estimates, the ONS use standard error estimates to calculate confidence intervals. Standard error percentages are used by the ONS; and they deem a percentage of over 25% to be unreliable.

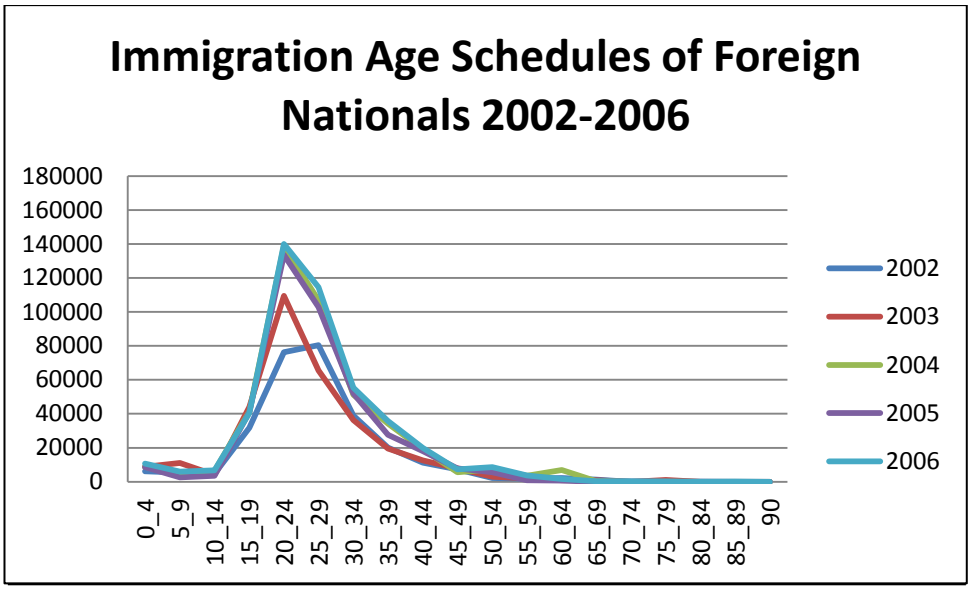

Figure 3.2: Immigration Age Schedules of Foreign Nationals, 2002 – 2006
Source: IPS

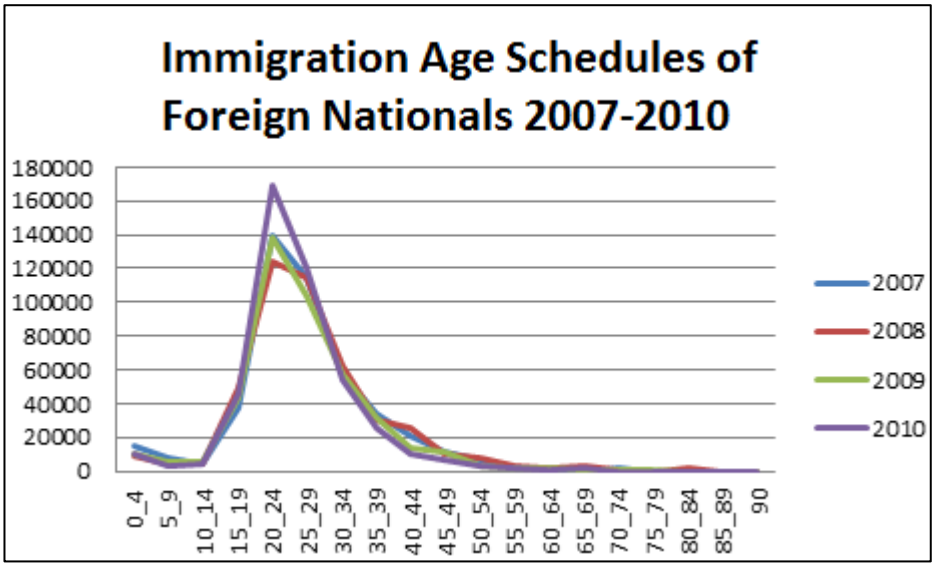

Figure 3.3: Immigration Age Schedules of Foreign Nationals, 2007 – 2010
Source: IPS

There are limitations to the standard error percentages for estimates of international migration taken from the IPS as they are designed for the IPS as a whole and not specifically for the subset of IPS migrants (ONS 2011 a). It is important to bear this in mind throughout and to view the standard error percentages and confidence intervals simply as an indication of reliability rather than a more precise estimate.

However, when one starts to look at flows which are based on a smaller part of the IPS sample, sampling variability in the age schedules becomes more apparent. Figure 3.4

shows the age schedule of immigration of German nationals, a significant migrant sender for 2002-2004. There is a large level of sample noise in this graph.

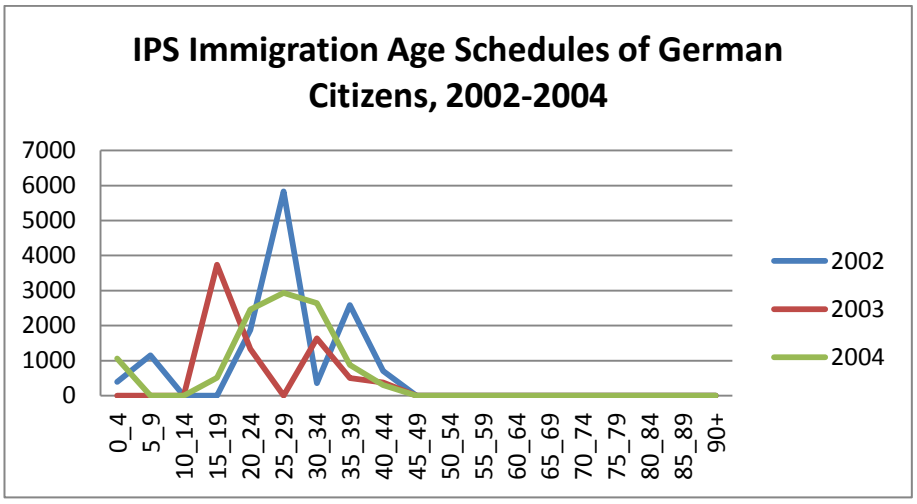

Figure 3.4: Immigration Age Schedules of German Citizens, 2002-2004
Source: IPS

The sampling variation in all but the total estimates and largest flows of immigration distorts true flow considerably. As outlined in chapter 1, it is this level of uncertainty in the main source data that the ONS use to estimate UK immigration, which is one of the main motivations for this research.

3.3.2 Non-IPS Components of Long Term International Migration

Each of the non-IPS components of the ONS estimate of LTIM are not available to use in this research and as such are not assessed in relation to true flow. They are included in the data audit, however, to give a complete picture of the main sources of data the ONS use to estimate long term international migration.

Home Office Asylum Seekers

The Home Office data on asylum seekers is an administrative data set and is therefore not designed to measure international migration. As such, the ONS makes assumptions about different categories of asylum seekers to establish how asylum seeker applications affect immigration and emigration flows respectively. The data consists of information on the principal applicant and their dependents and numbers on different types of asylum applicants – those who applied for asylum, were refused asylum, who have

appealed their asylum decision, who were returned home and those who withdrew (ONS 2011 a).

Through having information on these different 'types' of asylum seekers the ONS can therefore determine who in the data sets should be counted in the LTIM estimate for immigration or emigration. For example, those who leave the UK within a year of applying for asylum are not included in the estimate. It is thought, by the ONS, that most asylum seekers are not captured by the IPS as they are, dependent on exactly when they claim asylum (either at the port or after they have arrived in the country), either escorted over the IPS counting line or are not likely to have indicated that their intended duration of stay will be over 12 months, as it is assumed that they will answer the IPS question about duration of stay in the same way as they would an immigration officer (ibid: page 11).

Visitor Switcher and Migrant Switchers

One of the main limitations of the IPS is that it is based on individuals' intentions, which can change. As such, two components of the estimate of LTIM include an adjustment for people who state that their intention is to visit the UK, however, they decide to stay on as a migrant and vice versa. These changes of intention are taken into account to make the estimate of LTIM more robust.

In order to capture visitor and migrant switchers appropriate questions were introduced to the IPS methodology in 2004. However, this is another possible source of bias, as it is not clear if people, whose intention change, will respond to the switchers question in a representative way. The response to the switchers question is not something that can be taken into account this research though, as this information has not been made available.

Republic of Ireland

The method of incorporating flows to and from the Republic of Ireland into the LTIM estimate has changed over time quite considerably (ONS 2011 c). Before 2008 the ONS used data from the Central Statistics Office in Ireland (CSO) to estimate migration flows between the Republic of Ireland and UK. It made sense to use the data from the CSO as none of the routes between Ireland and the UK were historically covered by the IPS.

From 1999, though, the ONS started to survey routes between Ireland and the UK to compare IPS and CSO data to try and establish which is the most reliable. In 2008 the ONS concluded the estimates from the IPS to be the more reliable and have started use them to calculate these flows (ONS 2011 c). It is thought that these changes to the methodology mean that the ONS was underestimating immigration by approximately 10,000 per year and over estimating emigration by 2000 on average (ibid: page 2). These figures need to be treated with caution though, as flows of this size tend to be subject to significant sampling error and should just be treated as an indication that the IPS, in general, tends to estimate immigration from Irish citizens slightly higher than the CSO estimates.

Northern Ireland

According to ONS documentation, prior to 2008, international migration to and from Northern Ireland was estimated by the IPS, but according to the ONS (2011 c) there were not any ports surveyed in Northern Ireland

3.4 Alternative Sources of International Migration Data

Data on international migration flows tend to be considerably more accessible than adequate documentation regarding underlying definition and of the data collection practices employed to gather them (Zlotnik 1987 page 926). This is certainly the case regarding the administrative data that is detailed in this chapter, which has not been collected for the purpose of estimating migration. The thrust of much of Zlotnik's work is to emphasise the importance of international migration data being based on the consistent definitions and concepts, to allow comparability, internationally.

There has been a comprehensive initial review of the use of administrative data to improve international immigration estimates at a subnational level (see Bijak 2010 a and Boden and Rees 2010). The main focus of Boden and Rees' and Bijak's reviews is how alternative data can be used to identify better subnational distributions of different migrant streams. With the estimation of citizenship-specific immigration flows the focus of this research, the assessment and evaluation of data sources will look to build on Boden and Rees' (2010) contribution by including a review of administrative data sources that can be used to produce model estimates. Alongside the assessment of the IPS in section 3.3.1, the

main contribution of this chapter is an assessment of alternative sources of data in relation to true flow, using the four data assessment criteria outline in section 3.2.3.

3.4.1 Higher Education Statistics Agency

Information for this section, unless otherwise stated, is taken from the meta-data document supplied by the Higher Education Statistics Agency (HESA 2012). From this point in the thesis this data will be referred to as HESA data. The country detail provided by the HESA data is the citizenship of non-UK domiciled students. HESA collects data every year from UK universities. The data collected is provided to UK government and Higher Education (HE) funding bodies to ensure regulation of the sector. The data requested from HESA, for use in this research, is the citizenship-specific number of non-UK domiciled students, in their first year of study, at publicly funded HE institutions. This data is assessed below in relation to true flow and in this assessment more details about the nature of the data are outlined.

Definition

The requested sub-population provided by HESA are students who study at publicly funded institutions of Higher Education who are non-UK domiciled students. This predominantly includes students who have come from abroad to specifically study in the UK. Students who are excluded from this sub-population include incoming visiting and exchange students, post-doctoral students and students who are registered in the UK but are primarily studying outside of the UK. With regard to the assessment of true flow, it seems that this data set could contain a large proportion of students, through the exclusion of short-term exchange and visiting students, who are going to be usually resident in the UK for at least 12 months. There will be students who do not stay in the UK for the full 12 months, such as Masters degree students who finish within 12 months and students who drop out within their first year; but, it seems reasonable to say that they are in the minority in this data set. Consequently, with regard to the assessment of definition, it would seem that this data set is a reasonable reflection of true flow.

The total population of students is split into "years of study", subtly different to the year of course that the student is on. For example, if a student needs to retake the first year of a three year course they will be on the 4th year of study when they finish. With this in

mind, and to gain a better representation of a flow of migrant students, rather than the overall stock value, data on students who are in year 1 of study was requested from HESA. Furthermore, this will ensure that individuals are not double counted.

Coverage

Generally, the Higher Education Statistics Authority (HESA) collect data on all students registered at publicly funded Higher Education (HE) and Further Education (that provide higher level qualifications) institutions. With regard to the coverage assessment criteria, this means that students who are studying at privately funded institutions, such as language schools are not included and students at Further Education colleges, who are not studying at the HE level, are also excluded from the data. Furthermore, the HESA data only provides information on student flows so there is complete under-coverage of non-students. Another source of under-coverage from the HESA data is for children, as the vast majority of enrolled on HE courses are over the age of eighteen.

Bias

There should be no bias in the HESA data as it is not reliant on the migrant registering or responding to a survey question. The migrant's institution should provide HESA with the number of non-UK domiciled students that they have registered to study.

Accuracy

Due to the provisions of the *Data Protection Act 1998* HESA has a policy to release tabulations of data that are designed to prevent disclosure of personal information that can be identified to belong to a specific individual. Consequently, numbers are rounded to the nearest multiple of five with the exception of 0,1, and 2 which are rounded to 0. This will have a minimal impact on the accuracy of the data.

Furthermore, the HESA data does accurately measure the number of non-UK domiciled students who study at publicly funded HE institutions; the only source of error, because it is registration data, will be administrative; for example, a loss of records or erroneous data entry.

3.4.2 National Insurance Number Registrations (DWP Data)

This data is taken from the Department for Work and Pensions (DWP) website and is publicly available[4]. From this point in the thesis it will be referred to as DWP data. The country detail provided by the DWP data is for the citizenship of the individuals in the data set. Seeing as though, like many other administrative sources of data, this data has not been collected with the purpose of measuring international migration it is vital to carefully consider who is likely to be included. There will be individuals in the data set that do not match the UN definition of an international migrant and, equally as important, there will be people who are not included in the data set who do qualify as an international migrant according to the UN definition.

The DWP data used in this research are a record of "National Insurance Numbers" (NINo) allocated to adult overseas nationals entering the UK. A NINo is generally required by any overseas national looking to work or claim benefits / tax credits in the UK, including the self-employed or students working part time (DWP 2010). People pay National Insurance (NI) to enable entitlement to certain state benefits. It is paid whether people are employed or if they are self-employed. Foreign nationals, who intend to work, obtain a NINo through interview with their local job centre.

A NINo is effectively an individual's own personal account number for the social security system (DWP 2010). The number makes sure that historic NI contributions and tax are paid properly and recorded on individuals personal records. It also acts as a reference number for the whole social security system. As such, people are only issued with one NINo. For migrants this is when they originally apply for a NINo for work or benefits purposes.

Definition

Specifically in relation to the definition assessment criteria, there will be a substantial number of short-term and seasonal migrants who register for a new NINo but then leave the country soon after. These migrants do not necessarily match true flow in terms of usual residence, as they may deem UK their temporary home during their period of work, and in terms of duration of stay.

[4] Access to the tabulation tool can be gained here: http://83.244.183.180/mgw/live/tabtool.html

Coverage

It is not the case that every international migrant will have a NINo. International students should not be allocated a NINo, unless they are working part time or are claiming any kind of benefit - such as child benefit. For non-UK domiciled students, there is therefore some overlap in coverage between the DWP and HESA data.

Only foreign nationals who can legally work in the UK will be able to successfully apply for a NINo. As such, people who solely engage in the informal labour market sector will not be included in the DWP data. People in work start paying NI contributions if they are an employee (or an employer) and are aged over 16. As such, only people who are aged over 16, who are involved in the UK social security system, in any way, have a NINo. As a result migrant children are excluded from the DWP data.

Bias

A consequence of only being allocated a NINo once, with regard to estimating true flow, is that a person may migrate to the UK for work at some point in the past, say 2002 for example, but then leave the country for a long period of time. However, if they then return, in 2007 for example, they will not be required to re-apply for a NINo and will therefore not be picked up in this particular data set. If could bias true flow if certain migrant citizenships are more likely to migrate in this circular pattern.

Accuracy

With regard to accuracy, the DWP data does accurately measure the number of new NINo registrations of foreign nationals. Like the HESA data, a potential source of error, because it is registration data, will be administrative. One other source of inaccuracy could be non-random variability in the number of registrations. For example, there could be a time lag between someone arriving in the country and registering for a NINo.

3.4.3 National Health Service Patient Register for England and Wales (Flag 4 Data)

In the National Health Service (NHS) Central Register, an individual is usually given an NHS number when they are born. However, on registration with a GP, if someone needs an NHS number then they cannot have been born in the UK and must have migrated to the UK at some point in their life (Raymer et al 2012). They are assigned a code, which is known by the ONS as a "Flag 4 identifier". From this point in the thesis

this patient register data is referred to as Flag 4 data. There is no country detail, or information on the origin or citizenship of the migrants, provided in this data set. As such, it can only be used to provide information on the total number of immigrants.

Definition

To register with a GP one has to be usually resident in the UK. However, there is no duration of stay criteria for usual residence in respect of GP registration to receive health care on the NHS. As such, there could be short term migrants included in the flag 4 data, who do not intend to be usually resident in the UK for the 12 month period to correspond with true flow.

Coverage

It is clear, because the data set is only for England and Wales, that there is not complete coverage of immigrants for the whole of the UK. Theoretically though, every immigrant who arrives in England and Wales and is usually resident here could be included in the flag 4 data. As such, for England and Wales the coverage of this source of data is reasonably good.

Bias

This data set is contingent on self- registration of people with their GPs; consequently, there may be a substantial undercount of immigrants in the flag 4 data. It is well documented that young men are less likely to register with their GPs, which means that there are certain groups who could be systematically excluded from the data set (Raymer et al 2012). International students, even though they are usually resident in the UK, may be less likely to register with their GP as they may still see their country of origin as the appropriate place to be registered with a doctor. Furthermore, one would expect a large proportion of student migrants to be quite young, aged between 18 and 24. As result, the Flag 4 data could undercount student migrants, in relation to true flow.

For most migrants there will be a lag between immigration and registration with a GP, with no way of telling how many of the registrations will take place in the same year as the migration. For example, young men might only register with their doctor when there is a need to. However, with regard to just the time lag element of the bias data assessment, if it is consistent over time, in that the same proportion of migrants delay their registration

for the same amount of time year-on-year, the pattern exhibited by the Flag 4 data will be a good reflection of true flow. It will just simply lag behind reality.

Accuracy

Similar to the other sources of administrative data outlined in sections 3.4.1 and 3.4.2, the Flag 4 data, in relation to the accuracy assessment criteria, is a good source of information; as it is an accurate measure of the number of foreign nationals when they first register with a GP.

One source of inaccuracy though, is that the Flag 4 indicator is not retained on a patients record when a migrant moves within the UK and registers with a new GP. The Flag 4 data base is generated from a snapshot of the patient record at a given point in the year (ONS 2007). If a person migrates to the UK, registers with a GP and then moves internally before the snapshot is taken, they will not be included in the Flag 4 database. This could lead to a possible undercount.

3.4.4 Census Data

The Special Migration Statistics (SMS) from the 2001 Census provide immigration flows for 2000/01. This data is transition data, and outlines who was resident abroad 12 months before Census date. The country detail provided by the Census data is the country of previous residence of the individuals in the data set. Strictly speaking, this is not consistent with the data assessed previously, which has country detail, where available, on the citizenship of the migrants. An obvious limitation of this data is that it is only directly relevant to the estimation of a flow in 2001. However, with the limited amount of information on UK immigration it is worth assessing every source of available data.

Definition

Due to the data being transition based it is not certain that everyone in the data set will have been resident in the UK for 12 months or will have stayed in the UK for a period of at least 12 months, in total, after the Census. To be counted in the Census it is a requirement to be usually resident; but, it is not necessarily the case that people in this source of data will match the UN definition of a long term international migration.

Coverage

With regard to coverage, the main limitation is that the Census only covers England and Wales. Thus migrants who are usually resident in Northern Ireland and Scotland will not be included. For England and Wales, though, the Census theoretically has complete coverage for all people who are usually resident in the UK.

Bias

Areas of migration are linked to areas of under-enumeration (ONS 2001). This is because a large proportion of the migrant population are young students who often live in halls of residence and private rental accommodation that is hard for enumerators to gain access to (ibid). With a high proportion of migrants being young (cf. Rogers and Castro 1981) and students this means there could be a level of non-response bias.

Furthermore, non-response bias could vary according to the country of citizenship of the migrants. For example, language is often a barrier to completing the Census form. So, citizenships where English is not a first language could be less likely to complete the Census form than English speaking migrants. As a rule of thumb, one would expect the bias distortion of true flow for the 2001 Census to lead to an undercount of UK immigration, at varying levels dependent on citizenship.

Accuracy

The Census data accurately records respondents who were resident overseas one year prior to Census day; and, as with the HESA data, the only inaccuracy will be administrative.

3.5 Summary of Data Assessment

Table 3.1 below provides a colour coded summary of the assessment of each of the data sources outlined in this chapter. Green shading indicates that for a given data assessment criteria the source of data is good reflection of true flow. Amber shading indicates a reasonable reflection of true flow and red indicates that the data is not a good reflection of true flow for that given criteria.

Data	Definition	Coverage	Bias	Accuracy
International Passenger Survey	Immigration question in survey is designed with UN definition in mind.	Covers all formal immigration. Possibly under-samples at regional airports and ports for certain citizenships. Possible undercount for certain citizenships.	Survey is intentions based. Potential bias in declaration of intention for duration of stay and reason for migration	Sampling variability - country specific immigration flows are not very accurate. Totals, however, are more reliable.
NINo Registration Data (DWP)	Does not match the UN definition. There is no information on duration of stay and usual residence. Probable overcount short term migrants registering for NINo	Captures people who work formally in UK and claim tax credits or welfare benefits. Does not capture British migrants, full time students and children under 16. Probable undercount	Only discernible bias comes from only counting circular and repeat migrants when they first register for a NINo. Possible undercount.	Accurately measures the number of new NINo registrations by citizenship; error will be administrative. Non-random error in time lag between arrival and registering.
Non-UK Domiciled Students Data (HESA)	Does not match UN definition exactly, but slightly better than NINo. Students in their first year of study. Includes students who drop out but excludes students on short courses. Majority will study for at least 1 year.	Just public HE students. Students at language schools, private universities and further education colleges are not included. Doesn't capture British students. Does not capture children and older people. Probable undercount.	There is no discernible bias.	Accurately measures the number of non-UK domiciled students by citizenship; error will be administrative.
NHS Patient Register Data (flag 4)	Does not match the UN definition, as there is no information on duration of stay.	Does not capture British migrants and only covers England and Wales. But, theoretically, everyone who is usually resident can register	Only people who register for GP are included. Young men often do not register. Students might only register when they need healthcare	Accurately measures the number of new patient registrations of foreign nationals; error will administrative. Snap shot collection means that patients moving internally are missed.
2001 Census Special Migration Statistics Data	Does not match the UN definition exactly, as it is transition data and the duration of stay is unknown. Usual residence is a good match of true flow.	Only includes migrants who are usually resident in England and Wales. Theoretically should have complete coverage for England and Wales. Probable undercount.	Non-response bias of Census form. In particular, hard to count groups such as students, young people, those with English as second language, for example. Probable undercount.	Accurately records people who answer migration question in survey. Error will be administrative.

Table 3.1: Summary of Data Assessment Criteria for Each Source of Data

3.6 Conclusion

This chapter has established the concept of 'true flow' of migration. This is the focus of the estimation in the following chapters. Importantly each of the data sets available are outlined and then assessed in relation to true flow. This assessment is important in the establishing the methods that will be used to improve estimates of immigration to the UK.

The detailed assessment of each of the sources of data is summarised in table 3.1. the aim of the estimation of UK immigration is to take into account the summary of the distortion of true flow outlined in the table. From the assessment and audit of data in this chapter we have a high-level understanding of how true flow is distorted and indications of our certainty about each of these distortions. Proposing methods that take this into account is the next step in the research.

With regard to the alternative sources of migration data, the most comprehensive sources, that cover the whole of the UK and have information on the country-detail of immigration, are the HESA and DWP data sets. As outlined in detail in section 3.3 the IPS is a sample survey and therefore has problems with sampling variability; but alongside this, the main alternative sources of immigration data – DWP and HESA – are accurate in what they measure, even though as outlined in section 3.4 this is a distortion of true flow, because of issues outlined with regard to coverage and definition. This broad assessment of the IPS, HESA and DWP data forms the basis for the starting point of the analysis in chapter 4.

Chapter 4 – Log-Linear Modelling of UK Immigration Data

4.1 Introduction

Models for estimating migration flows are necessary because the data are often inadequate or missing (Raymer 2007). Aside from collecting better data through a population register, for example, there is a clear need to apply a methodology that makes the best use of the data available. Building on the conclusions of chapter 2 in relation to the difficulties of using migration theory in modelling of UK data and subsequent audit and assessment of data made in chapter 3, the focus of this chapter is to develop a data-driven model that combines sources of data in a statistical model.

As outlined in Chapter 3, there are definitional and conceptual differences between the sources of data available which lead to a distortion of true flow. Furthermore each source of data is of varying quality with regard to how closely it represents the true flow of UK immigration. Consequently, the models used in this chapter are selected to mitigate these problems as far as possible.

Statistical models can be used to estimate migration and, importantly with regard to this research, to combine data from different sources (Willekens 1999). There is, a growing literature on improving estimates of migration in migration flow tables (cf Abel 2010) and suggestions about how to deal with missing and unreliable data, mainly with regard to internal migration (Raymer et al 2006). The previous work that this research is based on is reviewed to provide the current state of the art. Statistical models, widely applied to similar research problems are log-linear models of contingency tables. Consequently, the method used in the second part of this chapter is an application of a series of log-linear models to combine multiple sources of data.

As a result, there is an account of contingency tables of categorical data and an introduction to how they can be analysed by log-linear models. This includes an outline of how log-linear models are commonly applied in statistical analysis. The log-linear model is built up in stages – starting with a main effects model and building up to two different models with separate offset terms. These two models are an application of an unsaturated

main effects model which, through the use of an offset, combine three separate sources of data to estimate student and non-student immigration flows to the UK in two separate models.

4.2 Review of Statistical Modelling of Migration

One can identify in the literature, longstanding statistical and modelling techniques that have been employed to help overcome the problem of missing or inadequate migration data. From estimating age profiles of migration, using the model migration schedules (for example, Rogers 1978) to estimating gross migration flows using spatial interaction models (for example, Stillwell 1978). Section 4.2.1 begins with a brief introduction to the modelling of spatial data and illustrates the link between spatial interaction models and log-linear models.

Following this there is a review of methods in section 4.2.2, which, primarily use log-linear models to firstly describe spatial patterns of migration and then are used to overcome problems with missing or inadequate data. Finally, in section 4.2.3 there is a review of the use of administrative sources of data in the UK to better estimate immigration to local authority areas. This section finds that there are various, established data-driven methods in the literature which look to overcome problems of insufficient, inadequate or incomplete data.

4.2.1 Modelling Spatial Data

The analysis of migration has a long history; stretching back to Ravenstein's (1885) seminal paper on empirically developing 'laws of migration'. In Ravenstein's paper, and much of the following literature on spatial interaction models, it has been argued that the pattern of the decline in frequency of trips could be parsimoniously represented by simple gravity model equations (Bennett and Haining 1985). These types of models, though, assume that people behave like particles (Willekens 1983) and in a uniform way.

The gravity model is often used to describe the interaction between pairs of geographical regions. Additionally, spatial data sets are often too large and complex to allow interpretation and need a formal way of being summarised that allows parsimonious inferences to be drawn. The gravity model provides one method, through the use of balancing terms, to describe spatial structures in large and complex data sets.

Bennett and Haining (1985) examine statistical approaches to the modelling of spatial structure and spatial interaction. In their evaluation of spatial modelling techniques, they set out two important families of models. The first type of model arises from the substance of geographical theory and represents a formalisation of that theory (explanatory model). The second type of model is developed as a description of a specific set of data (descriptive models). Such description may provide useful insights into the nature of the underlying processes or it may provide the basis for forecasting (Bennett and Haining 1985: page 1).

The general aim of using a theory driven approach is to borrow strength from covariate information that gives an indication of the levels of migration. Stillwell (2005) suggests that there are broadly two separate influences that need to be taken into account, a distinction between "those characteristics of individuals or households that are indicative of higher or lower propensities to migrate and those factors that actually determine whether a move takes place and which destination is selected." (page 5). Model fitting is then a part of theory testing. As such this form of data analysis is most commonly encountered where there is some prior knowledge about the underlying process of migration. As outlined in section 2.3, there is not a comprehensive literature on the theoretical determinants of international migration, so being able to disentangle the two broad considerations, outlined by Stillwell (ibid), is not possible.

There has been use of covariate information drawn from international migration theory in estimating international migration flow tables (Abel 2010). Data was used on various economic determinants of migration – trade and GDP for example – and covariates that took into account geographical influences were also used – distance and contiguity (ibid). As previously mentioned, UK data on actual migrations is poor enough; and it is severely lacking for the drivers and theories that directly explain movement of migrants (Stillwell 2005).

As mentioned in chapter 2, theory driven models have been used where data is missing (Raymer et al 2013). However, the main focus of the thesis is to make use of all publicly available evidence on UK immigration and to better understand the inherent uncertainty in these sources of data. A theoretical model is more suited to situations where parts of the data are missing, or where data is more reliable and explanations are being sought for the varying levels of migration evident.

The second family of models outlined by Bennett and Haining (1985) are descriptive models. Log-linear models have been found to be an important subsection of the spatial interaction models, which can be used to describe patterns of migration. Willekens (1983) advocates the use of a log-linear model in the study of spatial interactions. His reasons are as follows; log-linear models are formally equivalent to spatial interaction models shown through the balancing factors of a gravity model coinciding with the parameter values of the log-linear model; they act as an aid to focus on the data structure within a whole data table, rather than focussing on individual elements. Furthermore, the log-linear model simplifies the estimation of spatial interaction flows (ibid: page 188).

Estimating a log-linear using maximum likelihood treats the problem of calibrating spatial interaction models as a problem of statistical inference (Willekens 1999). Formally, this means that the maximum likelihood function of the log-linear model represents the likelihood that observations are predicted by the model, given the data (ibid: page 242). With regard to this research, this is an important consideration when considering model-fit. It has already been mentioned that sources of migration data are often unreliable and inconsistent. If one was to estimate a log-linear model of noisy data and the model fitted the data well, it would simply be describing the sampling variability within the data. Good model fit of noisy data does not necessarily mean that a good estimate of true flow has been made.

Log-linear models of contingency tables have been widely used in the analysis of migration flow tables (cf. Raymer 2007). However, as mentioned above, a log-linear model is not the only way to model data of this kind. A Poisson model with either row or column dummy variables are equivalent to log-linear regression models (Abel 2013). However, the use of log-linear models in the indirect estimation of migration flow tables are part of an established literature (ibid) and as such the remainder of this section will review the application of log-linear models in the estimation of migration flows.

4.2.2 Applications of Log-linear Models to Estimate Migration

An often-used approach in the migration estimation literature is the unsaturated log-linear model, used to smooth age and spatial structures in migration flow tables (Rogers et al 2010). Here, the marginal totals are used to impose a higher order structure on the cell values that could be deemed either irregular or unreliable. The notion of age structure is a

central concept in demography; however, there has been less of a focus on the spatial structure of migration in the discipline (Rogers et al 2002). There is an established literature in demography that uses mathematical expressions to describe the age patterns of migration (cf. Rogers and Castro 1981), whereas the spatial structure of migration is less well researched.

Using a log-linear model to analyse the spatial structure of migration was first suggested by Willekens (1983). Rogers et al (2002) describe a log-linear model which attempts to identify the relative push-factor at origin and pull- factor at destination and to express the origin and destination specific levels of spatial interaction in a migration flow matrix (page 30). The level of spatial interaction of pairs of origins and destinations is determined through the analysis of model parameters (ibid). However, in this context, the log-linear model employed by Rogers et al is suitable for analysing a multi-regional system and not necessarily a migration system that only includes one destination but many origins. Furthermore, the log-linear model is being applied in a situation where the aim of analysis is to explore the relationship between origins and destinations in terms of migration, rather than to estimate migration.

The aforementioned reviews of log-linear modelling of migration data mainly address the description and analysis of spatial interactions of migration flow matrices. Applying these methods verbatim, in the context of this thesis, would not be appropriate, as the aim is to apply methods that look to mitigate the effect of unreliable data on immigration estimates. A statistical model has to go beyond description to improve estimates of immigration to the UK.

As such, it is appropriate, at this point, to turn to Willekens' (1999) paper, which outlines various possible responses to missing migration data. One of the suggestions is to collect better data in place of the missing – in this case unreliable – data, which is beyond the scope of this thesis. A further suggestion, which does not fully address the main aim of the research, is to accept the incompleteness of the data and to minimise the distortions caused by the missing information.

The main aim of this research is use all of the publicly available data and evidence to estimate UK immigration. As such accepting the incompleteness of the IPS data is not an appropriate approach for this research. Willekens' (1999) final suggestion, however, is to estimate the missing (unreliable) data from available data, through developing a probability

model, where parameters are estimated using a combination of the incomplete data augmented by ancillary data (page 247). This is a more appropriate estimation strategy and Willekens' general approach to migration estimation is drawn upon in this chapter.

Following the development of the Poisson model for count data, Knudsen (1992) outlines how one can include auxiliary information into a model of count data. This is implemented through the use of an offset term. With respect to estimating migration, where the data is often missing or inadequate, the basic strategy in producing more detailed estimates is to use as much information as possible on the migration patterns that are being estimated (Raymer et al 2011 a); the offset term is one way to include this extra information.

In their application of a log-linear model to estimate international migration, Raymer et al (2011 a) assume that the marginal totals are known or have already been estimated. Their aim is produce more detailed estimates of internal migration in the UK through the combination of different sources of data, in a log-linear model with offset term (ibid). For example, the NHS data GP registration, which they use, provides relatively reliable information on the level of internal migration; however, there is a lack of information in that data set, other than the origin, destination, age and sex of the migrant (ibid).

Details on the ethnic groups of migrants is taken from 1991 and 2001 census data, and it is through the use of a log-linear model with offset term that the detailed year-on-year internal migration of the NHS registration data is combined with the detailed data on the ethnicity of migrants. Effectively, Raymer et al (2011) are using a statistical model to combine together sources of information, which have different relative strengths. Additionally, their model is not limited to description of a migration flow matrix, with the fitted values being the main results of interest. A model of this kind can be an appropriate way to combine different sources of information

Statistical models are one broad approach that can be used to better estimate international migration through the inclusion of auxiliary information. This auxiliary information can take the form of data that acts as alternative measure of international migration such as administrative sources, or covariate information which helps us to estimate migration more accurately. Having already ruled out the use of covariate information in a theoretical model, it is clear that methods, which have made use of

administrative and supplementary sources of data, are the most appropriate for this research.

4.2.3 Use of Administrative Data to Estimate UK Immigration at a Local Level

Since 2010 the ONS has used administrative data to help estimate the level of immigration at the local authority level based on a method developed by Boden and Rees (2010). Prior to this, data from the Labour Force Survey (LFS) was used. The total level of immigration to the UK was estimated by the IPS, and then estimates of where the IPS-estimated total resides at a local level were made using LFS and 2001 Census. The reason for using the LFS and Census to distribute an estimate of migrants to the local authority level is that the IPS response regarding where an immigrant intended to reside is biased towards major cities, especially London. This was shown to be the case when the usual residence of migrants was compared between the IPS, LFS and 2001 Census data. A further motivation for using the IPS, Census and LFS was that there is relative consistency regarding the duration of stay, as each source of data corresponds to the UN definition of an international migrant.

However limitations of using the LFS and the Census were identified – timeliness of the census during a period of high levels of immigration from EU accession countries and the exclusion of people aged under 16 in the LFS, for example (Bijak 2010). To address the limitations it was proposed that various sources of administrative data were utilised (ONS 2011 d).

Boden and Rees develop a model that estimates the geographical distribution of the total flow of international migrants, in the UK, at local authority level. Their work looks to circumvent the conceptual differences between data sets through the use of proportional distributions, rather than absolute migrant counts, in the estimation process (page 709). The administrative data that Boden and Rees suggested to use include National Insurance Number registrations, counts of overseas students, GP registrations and Census data (see chapter 3 for a detailed discussion of each of these data sources).

As outlined in chapter 3, there is not consistency in the definition of who can be identified as a migrant in each of these sources of data. The aforementioned use of

proportional distributions, though, meant the lack of consistency in definitions in each of the data sets was taken into account, as the estimates are constrained to IPS totals. Through pegging the proportional distributions to the IPS total, which in theory scaled up or down the local authority estimates, taken from the administrative data, a local authority estimate was made which satisfies the UN definition of an international migrant.

As previously mentioned, this formed the basis of the ONS' method for estimating immigration to local authorities across the UK. Importantly, though, the use of combining various sources of data at an aggregate level effectively increased the local geographical coverage in the IPS, whilst still satisfying the chosen working definition of immigration. Combining data at an aggregate level, where the IPS is used to constrain estimates to totals which satisfy the definition data assessment criteria, to mitigate the distortion to true flow is the general approach of the analysis in the remainder of this chapter.

4.3 Analysis of Immigration as Categorical Data

Chapter 3 outlines all of the sources of data that are available to estimate country specific UK immigration over time. With a data-driven approach being favoured, the statistical analysis that is conducted in this research predominantly uses categorical data. In this section there is an introduction to how categorical data, in this case immigration data, classified by citizenship and year, can be analysed both using descriptive statistics, in the form of a contingency table in section 4.3.1; and using inferential statistics, in the form of log-linear models of contingency tables in section 4.3.2.

The specific nature of the data used in this chapter, including elements of the data assessment carried out in chapter 3, is linked to both the contingency tables that the data is arranged in and the log-linear models applied. Finally, there is a justification for the application of log-linear models in the estimation of immigration, rather than their standard use of describing higher order structures, associations and interactions between categorical variables in contingency tables.

4.3.1 Contingency Tables

One of the overall aims of this research is to make use of all available evidence to estimate citizenship specific immigration flows to the UK. All the available data, outlined in chapter 3, is categorical. Furthermore with the analysis being data-driven (rather than

theory driven), the only categories of interest in this research are country of citizenship and time in years, which can be displayed in a contingency table.

These two categorical variables, with country of citizenship for the sake of description here denoted by X with *i* countries and year denoted by Y with *t* years. A table of this kind displays the number of observations at combinations of possible outcomes for the two variables (Agresti and Finlay 2009). This means, in a contingency table of immigration by country of citizenship and time there are *i* times *t* cells to be estimated. The cells contain frequency counts of immigration cross classified by country of citizenship and year. This takes the form of a contingency table with *i* rows and *t* columns and is called an *i*-by-*t* table (cf. Agresti 2002).

From chapter 3, it is clear that larger flows of immigration estimated from the IPS are deemed more reliable as they are less subject to sampling error. If the IPS data is arranged into a two way contingency table, then the margins – the total immigration of each citizenship over the period 2002 – 2010 and the total immigration of each year over all countries – would be subject to less sampling variability, due to these estimates being made up of a larger part of the overall sample than estimates from specific cells within the table.

With the data arranged in a contingency table, an appropriate model to constrain the *i* times *t* cells to the more stable margins, is a log-linear model. With a two way table comprised variables X and Y, one can rearrange a log-linear model to be expressed as a multinomial model and vice versa. Multinomial models can include a continuous predictor as an independent variable; however, as outlined in chapter 2, the statistical models in this thesis are data-driven and do not include covariate information taken from migration theories.

Log-linear models of contingency tables can only be applied to independent categorical variables. Specifying and computing unsaturated models is also straight forward. With the current state of the art of standard models of migration outlined in section 4.2, including methods which constrain administrative sources to the IPS and the straight forward nature of fitting unsaturated models, an appropriate first method to estimate UK immigration is a log-linear model of contingency tables. Models of independence and unsaturated models are outlined in more detail in the following section.

4.3.2 Log-linear Models of Contingency Tables

This section provides an introduction to log-linear models of contingency tables. Firstly there is a brief outline of the log-linear model as a part of the generalised linear model family. Following this there is a description of the key component parts of fitting a log-linear model – the assumptions the model is based on, how the observed data has been transformed so that one can fit a linear model, assessing the fit of the model and offset terms.

As mentioned in section 4.3.1, the main focus of this chapter is fitting log-linear models estimate UK immigration, through combining sources of data at a macro level in a way which mitigates the distortions of true flow as much as possible. Referring to the summary table 3.1 of the data assessment in chapter 3 the ultimate aim of fitting a log-linear model is to combine different sources of data to in a way that has close a reflection of true flow for each of the four data assessment criteria. As such, having a background understanding of how log-linear models of contingency tables are estimated is vital. This is outlined below.

The log-linear model is part of the Generalised linear models (GLM) family. GLMs extend ordinary regression models, allowing one to model non-normal response distributions and modelling functions of the mean (Agresti 2002). GLMs customarily include both random and systematic components (Nelder and Wedderburn 1972) and a link function (Agresti 2002). The random component of a GLM is the response variable, which is made up of independent or correlated observations and is from a distribution in the natural exponential family (ibid). The natural exponential family is a class of probability distributions that includes many common distributions used in the computation of a GLM – the normal, Poisson, gamma, binomial and negative binomial distributions. The response variable is seen as random as it could vary if the sample or population of study changes. The systematic component relates to the combination of explanatory variables called the linear predictor (Agresti 2002). Finally, **t**he link function connects the random and systematic components of the GLM. The link function allows the mean to be non-linearly related to the predictors (ibid).

Log-linear models are normally used to help describe association patterns among a set of categorical response variables of count data (ibid). The log-linear model is a GLM that assumes a Poisson distribution and uses the log link function (Agresti 2002). The

Poisson distribution is unimodal and has a single parameter greater than zero, which is both its mean and variance. The parameter of the model is the expected number of events, in this case the number of migrants. As it is impossible to have a negative number of events, the log transformation is used when relating the parameter to the variables and the data (Willekens 1999).

Log-linear models are mainly of use where there are at least two variables, in a contingency table, that are responses (Agresti 2002). This is the case, for example, when one models immigration cross tabulated by year and origin. A basic log-linear model of independence takes the following form,

$$log(n_{it}) = \lambda + \lambda_i^X + \lambda_t^Y \quad (4.1)$$

where λ is the overall effect of the whole sample, λ_i^X is a row effect and λ_t^Y is a column effect, also known as main effects. The larger the value of λ_i^X the larger the value of the expected frequencies predicted for row *i*. This is also the case for the expected values in column *t*, in respect of λ_t^Y.

For the independence model one of the λ_i^X and λ_t^Y terms are redundant, with most software, dependent on the constraints specified, setting the parameter for the last category to equal zero. During the interpretation of log-linear model parameters this needs to be taken into account. However, at this stage, with the focus on estimation rather than description or explanation of the drivers of immigration, there is not necessarily a need to focus on the interpretation of parameters. Accordingly, the results sections in this chapter will mainly analyse the fitted values rather than focussing on interpretation of model parameters.

Sometimes it is necessary to express a log-linear model so that one of the explanatory variables has a known co-efficient. A variable with a known coefficient, in a log-linear model is known as an offset term, where the offset coefficient is equal to one. Consequently, this term in the model is fixed. Often models that include an offset are applied where there have been observations of events over varying length of times. Thus, the offset is used as a term to indicate the level of exposure to the event of interest. Where time is auxiliary information in the example given above, there is no reason why, for the purpose of this research, an offset term cannot be used to introduce auxiliary data sets to a log-linear model.

The fitted values from the model are also the expected frequencies for the X^2 and G^2 tests of independence, which double up as the goodness of fit statistics for log-linear models (Agresti 2002: page 205). For example, a large X^2 or G^2 statistic with a small p-value, means that the null hypothesis is rejected and that there is significant difference between the observed and expected values; thus, indicating that the model does not fit the data well.

Assessment of model fit for log-linear models is very important in circumstances when one is using a log-linear model to describe and explain relationships between variables. This would be the case in an application where log-linear models are fitted to identify the structures and associations in a contingency table. In relation to this research though, models of data that have large sampling error that fit well are not necessarily an indication that a model is performing well in respect of estimating true flow of immigration. One's aim, in the context of noisy unreliable data, is not to over-fit the model; rather, to find ways of combining reliable elements of different sources of data. However, the extent that changing the specification of the models affects model fit in this chapter is considered.

The main characteristics of the log-linear framework outlined about – constraints to marginal effects and the ability to include auxiliary information via an offset term – are the main motivating factors for the model applications in section 4.4.

4.4 Log-linear Model of UK Immigration

This section details the development of log-linear modelling of UK immigration in this chapter, starting with a main effects model building up to two log-linear models with offset terms of student and non-student immigration. Firstly in section 4.4.1 a main effects model of IPS data is fitted. This model shows the how the main effects in the model translate into fitted values and how the model is constrained to the marginal totals.

In sections 4.4.2 and 4.4.3 log-linear models with administrative data offset terms are applying to the IPS, HESA and DWP data. The models specified in section 4.4.3 form the basis of a recommendation for the ONS. Building on the suggestions of Raymer et al (2011 a) this approach should help fulfil the Eurostat requirements for international migration data in the estimation of citizenship-specific migration flows.

Before the specification of the models, though, it is necessary to consider in detail some of the assumptions of log-linear models. Inferential statistics is concerned with drawing conclusions about quantities that vary. With regard to the data available to this research including a consideration of the natural variation from the sample based IPS data is problematic. The IPS data available has already been weighted up to the population level and the survey weights are not available. As such, the inflated population numbers are used in the models. This makes statistical inference problematic, as the standard errors and confidence intervals of the parameters cannot be estimated, for example. Furthermore, confidence intervals of the fitted values also cannot be estimated. Consequently, obtaining an estimate of the uncertainty of the log-linear estimates of immigration is not possible.

A further assumption of log-linear models of contingency tables is that each of the observations is independent. In a two way table this means that the probability of any given column response is the same in each row, and vice versa (Agresti 2002). To test whether this is the case, a Chi-squared test of independence is carried out on the IPS data. With a very large X^2 of 537,880 there is strong evidence against independence. Immigration not being independent does make intuitive sense from what we know about perpetuation of migration flows outlined in chapter 2.3.1. With the emphasis of this chapter being on using a log-linear model as tool to combine sources of UK immigration data, the lack of independence in the data is not too much of a problem. The purpose of the model is to simply constrain estimates to the margins of the IPS. Making statistical inferences from the parameters of the model is not the primary concern of this analysis.

As documented throughout this chapter the data is arranged in an *i*-by-*t* contingency table, where *i* denotes the country of citizenship and *t* denotes year. Another assumption of the log-linear models that are fitted to the contingency tables of migration is that the quality of the data being used does not depend on either of these two categorical variables.

4.4.1 Main Effects Model

The starting point of the log-linear analysis is a main effects model of the IPS data. As mentioned in section 4.3.2 a main effects model has a row and column effect. Applying

the same approach as Rogers et al (2010), this model uses the marginal totals to impose a higher order cell structure on data that is deemed irregular and unreliable. A model of immigration by country of origin over the time period 2002 – 2010 takes the form

$$log(n_{it}) = \lambda + \lambda_i^{IPS} + \lambda_t^{IPS} \quad (4.2)$$

where $log(n_{it})$ is the log-transformed estimated value of immigration, λ_i^{IPS} is the country of citizenship effect for row *i* and λ_t^{IPS} is the time effect for column *t*, and superscript *IPS* denotes that this is a model of IPS data. The model specified by equation 4.2 is an unsaturated log-linear model, as it does not include the interaction term of the origin and year effects. A model that included an interaction effect, a saturated model, would explain the observed data completely.

This model, because it is imposing the higher order structures from the marginal totals, assumes for each individual country of origin, the total IPS estimate of immigration over the time period 2002 – 2010 is reliable. Furthermore it assumes that the IPS estimate of total immigration for each individual year is also reliable. This informs the selection of data that is used in the model. The ONS provides estimates of 'standard error percentages' (see chapter 3.5.1 for discussion); but here, they are simply used as a guide to data selection, and are not used to estimate confidence intervals. From the standard error percentages provided on a data set of 2002-2010, it appears that the top 30 citizenship flows of immigration to the UK, when totalled over the time period, have acceptable standard errors, according to the ONS. The totals of immigration for each year respectively, as you would expect, also have acceptable standard error percentages.

Consequently, the main effects model of the top 30 flows (plus an 'all other' category) for the years 2002-2010 is estimated. This is a model of a 31-by-9 cell contingency table. The degrees of freedom (*df*) in this model equal the number of cell counts minus the number of model parameters; so, for a saturated model, there are no degrees of freedom. For the independence model, equation 4.2 outlined above, for a 31-by-9 contingency table, there are $df = 279 - (i-1) - (t-1) - 1$, which is equal to 240 degrees of freedom. This, coupled with the significant levels of sampling variability exhibited in the vast majority of immigration flows over time, one would expect that, according to the likelihood ratio test of goodness of fit, the model does not fit the data well. This seems to be the case - the statistic used to indicate model fit, the likelihood ratio (G^2) and the p-

value are shown below in table 4.1. The large value of G^2 and the small p-value indicates that the model does not fit the data well.

Table 4.1: Likelihood Ratio Statistic for Model 4.2, Main Effects Model.

Model	G^2	P-Value
4.2, Main Effects	569313	<0.001

Figure 4.1 below shows the results of the top 5 flows for the model specified by equation 4.2. The main effects model, equation 4.2, produces results that do not necessarily improve estimates of immigration to the UK. Effectively, due to the marginal constraints, the patterns exhibited in the results are simply a reflection of the total pattern of immigration from all countries. This assumes that immigration from individual countries has exactly the same pattern as the total level of immigration. It is apparent from these results that the λ_t^{IPS} term in the model sets the pattern over time exhibited below and that the λ_i^{IPS} term places the citizenships in order from the largest to the smallest flow.

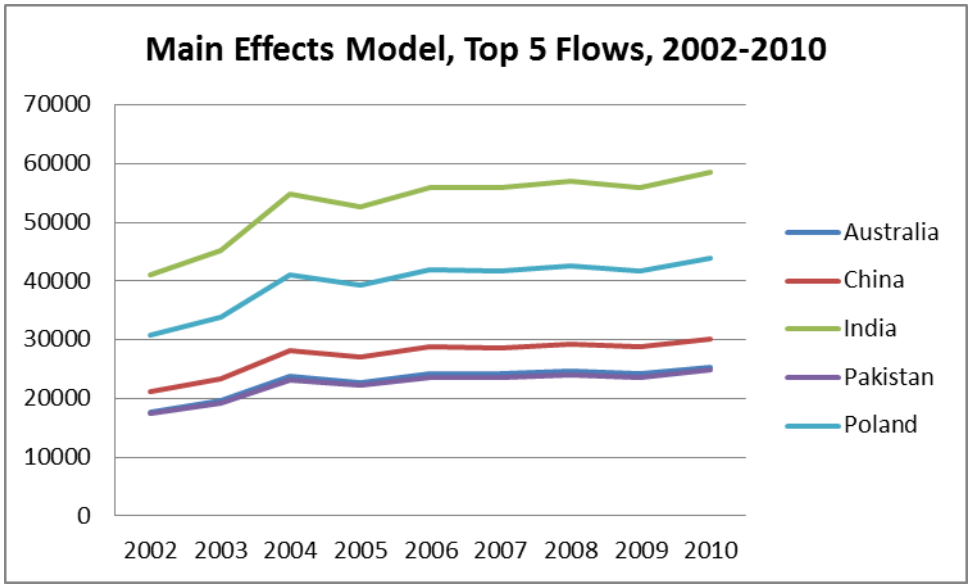

Figure 4.1: IPS Main Effects Model of Top 5 Citizenship flows, 2002 – 2010

If one were to plot all 31 flows on one graph, none of the respective lines showing the fitted values would cross, due to the constraints of the main effects terms in the model. From the review of UK immigration in chapter 2 and the general descriptive analysis in chapter 3, we know that country-specific patterns of immigration are heterogeneous. A main effects log-linear model, such as the model specified above, does not allow for this

heterogeneity and its fitted values are not a good representation of true flow over time. Consequently, further auxiliary information is required, through the use of an offset term.

4.4.2 Main Effects Model with Offset Term

As mentioned in sections 4.2 and 4.3, one way of incorporating auxiliary data into a log-linear model is through the use of an offset term. In chapter 3, it became clear that the only other sources of data that have full coverage of the UK, with variables that can be used to indicate country of citizenship over time, are data on non-UK domiciled students and NINo registrations from the HESA and DWP data sets, respectively. Consequently, the log-linear models with offset terms in this chapter make use of the two aforementioned auxiliary data sets.

Effectively an offset model is similar to the independence model specified by equation 4.2, whereby the model is constrained to the marginal row and column effects. The key difference, though, is that the offset term allows the citizenship-specific patterns of immigration to vary according to the patterns exhibited in the auxiliary data set. For example, if the student HESA data is chosen as the offset term, one is effectively saying that the pattern of student immigration to the UK, for a given country, is the same as the pattern of the true flow of immigration; with the fitted values being constrained to the reliable part of the IPS data.

Figure 4.2 below gives a colour coded illustration of a log-linear model with offset term and equation 4.3 gives an example of a log-linear model with offset term. The green-shaded areas denote parts of the model that are deemed to be a good representation of true flow and red-shaded areas where true flow is distorted by the data collection process. As previously outlined the main effects of the model, denoted by λ_i^{IPS} and λ_t^{IPS}, take into account the marginal distributions of the variables and are consequently based on a larger part of the overall sample than individual cells within the contingency table.

As such the margins of the diagram, representing the main effects of the model, are shaded green and the cells within the contingency table, representing the cells of the IPS *i t* contingency table, that are subject to large amounts of sampling variability, are shaded red. The offset term, $log\left(n_{it}^{offset}\right)$ is shaded green; a reflection of how each of the auxiliary data sets have a high level of accuracy. Both are administrative sources of data; they are not

subject to sampling error and are accurate representations of the number of non-UK domiciled students and the number of overseas nationals who register for a NINo.

Consequently, figure 4.2 shows how the reliable patterns exhibited in the auxiliary data sets are constrained to the reliable IPS marginal totals. Specifying a log-linear model with an offset term, using administrative data in this way, pieces together an estimate of immigration which starts to use the respective strengths of each source of data used, with regard to the data assessment criteria. Referring back to the summary of the data assessment criteria of each data set in table 3.1, an ideal estimate as result of the model specified, be shaded green for each of the assessment criteria.

The main effects parameters utilise the strength of the IPS data's close match to the definition and coverage criteria without being subject to its weakness in relation to the accuracy criteria, as a result of sampling variation. The offset term utilises the accuracy of administrative sources of data, but through the marginal IPS constraints, is not subject to not matching true flow with regard to definition and coverage. The remaining assessment criteria, bias, however, cannot be taken into account in this framework.

$$log(n_{it}) = \lambda + \lambda_i^{IPS} + \lambda_t^{IPS} + log(n_{it}^{offset}) \quad (4.3)$$

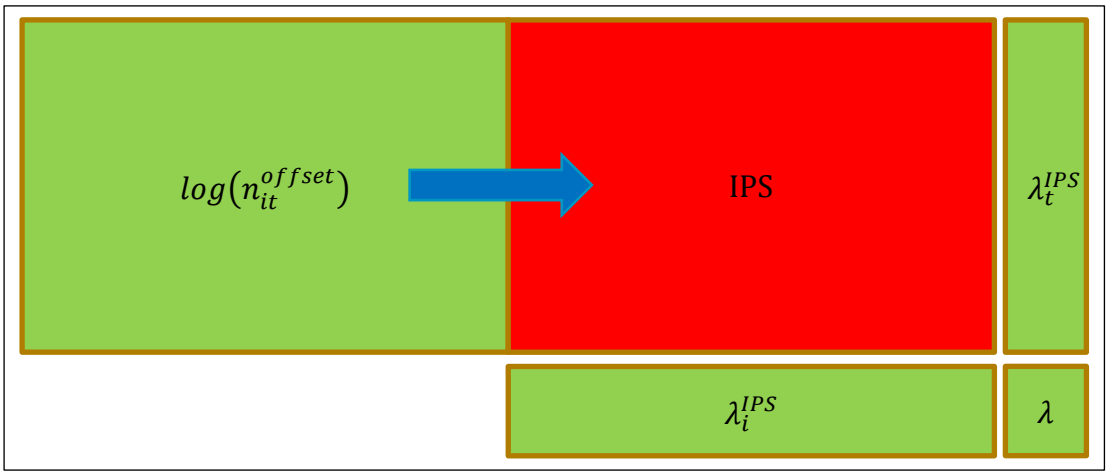

Figure 4.2: Colour Coded Diagram Illustrating Aims of Log-linear Model with Offset Term

Equations 4.3 and 4.4 specify two log-linear models of independence with HESA and DWP offset terms respectively. They take the same form as the main effects model – equation 4.2 - with the addition of the offset term $log(n_{it}^{HESA})$ where the auxiliary data

being used is the HESA data on non-UK domiciled students and $log(n_{it}^{DWP})$ where the auxiliary data being used is the DWP NINo registration data.

$$log(n_{it}) = \lambda + \lambda_i^{IPS} + \lambda_t^{IPS} + log(n_{it}^{HESA}) \quad (4.4)$$

$$log(n_{it}^{*}) = \lambda + \lambda_i^{IPS} + \lambda_t^{IPS} + log(n_{it}^{DWP}). \quad (4.5)$$

The superscripts DWP and HESA simply denote which auxiliary data set is being used, as an offset term in each respective model. For clarity here, n_{it} is the estimate of immigration for students using the HESA offset term and n_{it}^{*} the estimate that uses the DWP offset for non-students. Each of these models have the same degrees of freedom as the main effects model, specified by equation 4.2.

Having fitted models 4.4 and 4.5 it is clear that they both have relatively poor model fit; with large likelihood ratio statistics and small p-values (see table 4.2). Interestingly, though, the G^2 statistic for equation 4.5, that uses the DWP NINo registration data as an offset, is considerably smaller than the model specified by equation 4.2 indicating better model fit. The model specified by equation 4.4, with the HESA data offset term, also fits slightly better than the main effects model. The improved and better fit of model 4.5, in comparison to models 4.2 and 4.4, suggests that the inclusion of the DWP auxiliary data produces fitted values that are a closer match to the observed IPS values of immigration. This suggests that there are more similarities in the patterns exhibited in the DWP data of new NINo registrations of foreign nationals and the IPS estimates than the student registration HESA data and the IPS.

Table 4.2: Likelihood Ratio Statistic for Model 4.2, the Main Effects Model, and Models 4.4 and 4.5 that use HESA and DWP Data as Offset Terms.

Model	G^2	**P-Value**
4.2, Main Effects	569313	<0.001
4.4, HESA Offset	417096	<0.001
4.5, DWP Offset	270586	<0.001

There is a clear limitation to estimating log-linear models with offset terms where the main effects are based on all IPS data, for the 31 flows. As outlined in chapter 3, both the HESA and DWP data do not cover all UK immigration as they systematically exclude

certain migrant groups. However, this is not taken into account if all 31 flows are modelled using a single offset term. The coverage of each offset term is incomplete; whereas, the IPS data has more complete coverage of immigration to the UK. This is illustrated below by figures 4.3 and 4.4 which show the estimated values of immigration from models 4.4 and 4.5, for Poland, respectively. As outlined in chapter 2, following the expansion of freedom of labour movement in 2004, Poland is predominantly a sender of migrant-workers to the UK.

For figures 4.3 and 4.4 the same scale has been used on the Y-axis to allow for comparison. As expected, from the literature outlined in chapter 2, the graphs show clearly, due to the considerably smaller figures in the HESA data, that Polish migrants are predominantly non-students. Furthermore they show that using offset terms in a model of IPS totals, where there is no separate consideration of students and workers in the marginal effects, produces estimates for non-student migration and student migration that are very similar. This certainly does not reflect the true flow of migration for Polish citizens, as the student flow is being overestimated.

Modelling the data in this way is also problematic when one considers a citizenship where there are comparable numbers of student and non-student migrants. China for example, has substantial numbers of new students and people registering for a NINo. However, if one uses DWP and HESA offsets separately on the whole of the IPS data then, similar to the student immigration of Polish citizens, one would be overestimating both the student and non-student immigration of Chinese citizens, if the IPS margins are deemed reliable.

If the models are to be constrained to the IPS totals for each of the 31 flows and 9 years then one can only apply a single offset term; rather than two offset terms applied to the same IPS data. Estimating a combined offset, say from an ordinary least squared regression to provide an estimate of the balance between the HESA and DWP data, to be used as a single offset term, is problematic. An offset term within a log-linear model is assumed to be fixed; whereas the fitted values of a regression model are assumed random, thus ruling out such an approach to estimation. Unfortunately, with regard to sample size, this means that in order to include auxiliary data the IPS needs to be split into student and non-student sub-samples, so that each offset term is applied to the appropriate part of the IPS data. This results in marginal constraints which are based on a smaller part of the IPS sample.

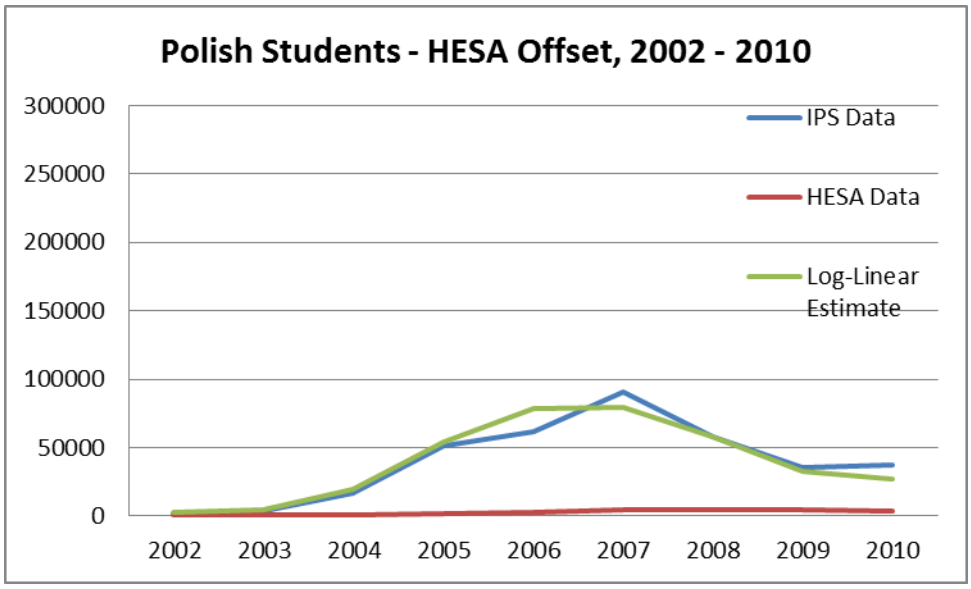

Figure 4.3: Comparison of IPS and HESA Data with Log-linear Estimate of Student Immigration from Equation 4.4 of Polish citizens, 2002-2010

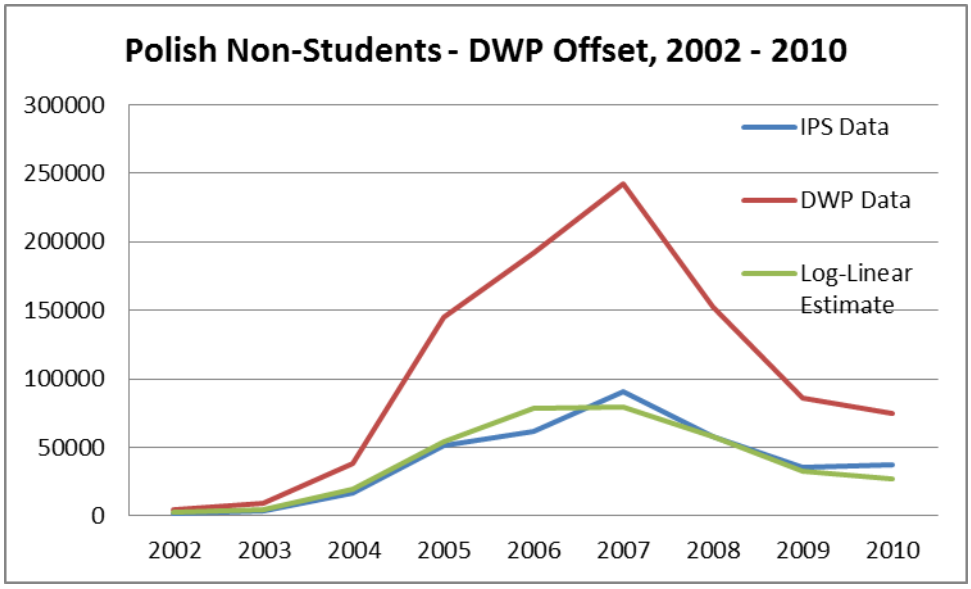

Figure 4.4: Comparison of IPS and DWP data with a Log-Linear Estimate of Non-Student Immigration from Equation 4.5 of Polish citizens, 2002-2010

Consequently the next step in estimation, within a log-linear framework, is to separate the IPS data in a way so that the HESA and DWP offset terms can be applied to the relevant sub sample.

4.4.3 Separate Student and Non-Student Models

The previous sections of analysis have made it clear that, in order to consider two different sources of auxiliary data, within a log-linear framework one has to specify two models, which estimate immigration with the HESA and DWP data as separate offset terms, respectively. As part of the IPS surveying process, there is a question on main reason for migration. This variable is the most effective way of splitting the IPS data in two, so that two separate log-linear models can be estimated. From chapter 3.6, through the consideration of how closely the data reflects true flow, it is difficult to have an exact idea of the types of migrants that make up the DWP data set. People do not necessarily register for a National Insurance Number just for work purposes.

Consequently, it is safe to assume that it is difficult to predict how people who are included in the DWP data set of new NINo registrations would respond, if sampled, to the question on primary purpose of migration in the IPS. One can envisage, however, that students are a group of migrants who are more likely to answer the question on reason for migration most predictably. This is based on the assumption that people, who are classified as non-UK domiciled students in the HESA data set are very likely, if sampled, to state that their primary purpose of migration is to study in the UK. With these assumptions in mind, the most appropriate way to split the data is into students and non-students.

The variables in the IPS that indicate the main reason for migration are as follows: accompany/join, definite job, formal study, looking for work, no reason stated, other. To split the data up, the estimates of formal study were placed in the student group and all the other categories, except 'no reason stated' were placed in the non-student group. The respondents who did not state a reason were assumed to have an equal probability of being students or non-students, so were split equally between the two.

The student and non-student models take the same form as the model with offset term specified by equations 4.4 and 4.5. The offset terms for the student and non-student models, respectively, are included as follows:

$$log\left(n_{it}^{S}\right) = \lambda^{IPS.S} + \lambda_i^{IPS.S} + \lambda_t^{IPS.S} + log\left(n_{it}^{HESA}\right) \tag{4.6}$$

$$log(n_{it}^{W}) = \lambda^{IPS.W} + \lambda_i^{IPS.W} + \lambda_t^{IPS.W} + log(n_{it}^{DWP}) \tag{4.7}$$

where n_{it}^{S} and n_{it}^{W} are the fitted values for students and non-students, and superscripts $IPS.S$ and $IPS.W$ denote the student and non-student separation in the IPS.

Both are log-linear independence models, with origin and year main effects and offset terms that are made up of the HESA and DWP data respectively. As outlined previously, in section 4.3.2, model fit and interpretation of parameters are not the main focus of this analysis. Estimation of true flow, rather than of the drivers of immigration, is the sole focus of the estimation in this chapter and the remainder of the section illustrates some main characteristics of the results.

As with the previous models, however, the likelihood ratio statistic G^2 is calculated (table 4.3). In addition to models 4.6 and 4.7, whose fitted values are used to estimate immigration, main effects models for both students and non-students are estimated. Main effects models, in section 4.4.1 are found to be an inappropriate way of estimating true flow; however they are estimated here simply to determine whether the addition of an offset term improves model fit. Similarly to the results for models 4.2 and 4.3, the inclusion of an offset term for models 4.6 and 4.7 improves model fit with smaller G^2 statistics

Model	G^2	**P-Value**
Student Main Effects	280138	<0.001
Student HESA Offset (Model 4.6)	246452	<0.001
Non-Student Main Effects	543796	<0.001
Non-Student DWP Offset (Model 4.7)	299822	<0.001

Table 4.3: Comparison of Likelihood Ratio Statistics for Non-Student and Student Main Effects Model and Models Specified By Equations 4.6 and 4.7 that have HESA data and DWP Data as Offset Terms

There are 62 sets of 2002-2010 flows estimated from the two models – the top 30 flows (plus an all other category) for both non-student and student flows. There is not room to display and analyse all of these estimates of true flow, so, specific flows of interest that have been identified. Examples of these will be selected and analysed below.

Firstly, as outlined in chapter 2, the expansion of freedom of labour movement within the EU led to a significant influx of migrants from Eastern Europe, and especially Poland, to the UK, with over half a million Polish born citizens resident in the UK according to the 2011 Census (ONS 2012 b). This flow is predominantly labour-based, with migrants coming to the UK to look for work. Chapter 3 concluded that the patterns exhibited in the IPS data for large flows, such as Poland, are relatively reliable. As such, one would expect the IPS non-student estimate to be relatively similar to the estimate of the non-student model. Figures 4.5 and 4.6 below show the results of the student and non-

student models from equations 4.6 and 4.7. Please note that for presentation purposes, due to the large difference in the estimates of students and non-students, the scales on the graphs are not comparable.

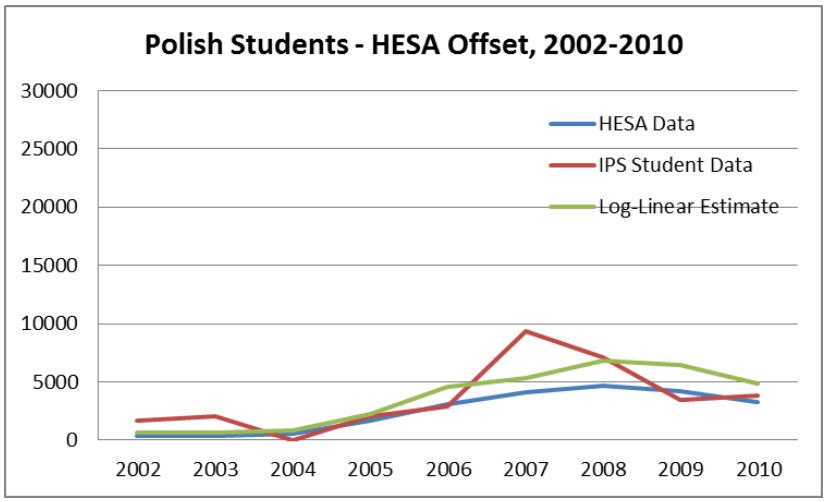

Figure 4.5: Comparison of IPS and HESA Data with Log-linear Estimate of Student Immigration, from Equation 4.6, of Polish citizens, 2002-2010

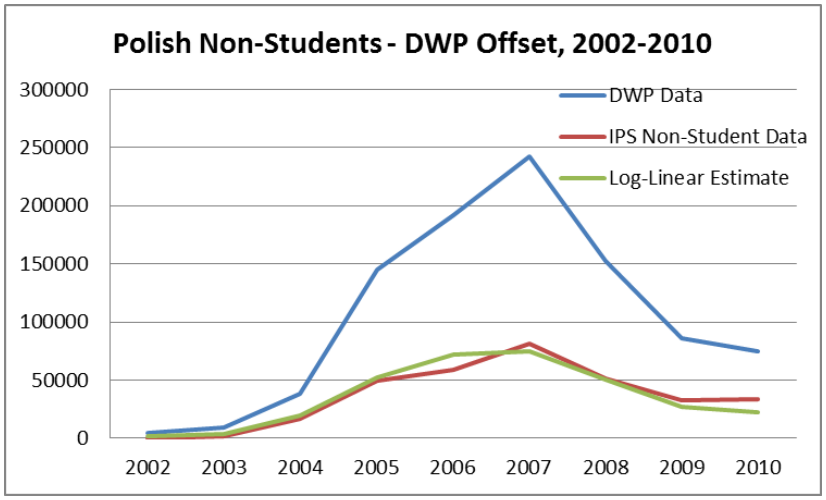

Figure 4.6: Comparison of IPS and DWP Data with Log-linear Estimate of Non-Student Immigration, from Equation 4.7, of Polish citizens, 2002-2010

From these results, one can see how the model is benchmarked to the marginal totals of the IPS. For example in figure 4.6 it is clear that the total in the DWP is far higher than the total of the IPS data. The IPS estimate of student migration from Poland is deemed less accurate than the non-student IPS estimate, as it is based a much smaller part of the overall migrant sample. The model, specified by equation 4.6, borrows the reliability

of the year on year variation of immigration of students from the HESA data set to produce a smoothed estimate bench marked to the marginal totals of the IPS; the $\lambda_i^{IPS.S}$ and $\lambda_t^{IPS.S}$ terms in the model.

It is also clear from figures 4.5 and 4.6 that modelling students and non-students separately, using the reason for migration variable to split the IPS data accordingly, has addressed the problems like the overestimating flows such as the student migration of Polish citizens, outlined in previous section.

The trend exhibited by the model result in figure 4.5 is taken from the pattern of Polish citizens from figures in the HESA data set. The pattern is smoothed and peak immigration is shifted from 2007 to 2008 in comparison to the estimate from the IPS data. This accords with the pattern exhibited in the HESA data.

To verify these patterns, one needs to refer back to chapter 2. We know from the literature that there was a rapid expansion of immigration from Poland following the expansion of freedom of movement within the EU; and, that this expansion was mainly driven by labour migration. This corroborates the model estimates; as the non-student model estimates a much larger immigration of Polish people in comparison to the student model.

However, one limitation of this model becomes apparent when one looks at the large difference between numbers in the DWP data (peaking at nearly 250,000) and the IPS non-student data (peaking at 80,000), against which the model results are benchmarked. The larger quantity exhibited in the DWP data could be explained by a distortion of true flow in relation to the 'definition' assessment criteria outlined in chapter 3.2 and detailed in relation to the DWP data in chapter 3.4.2. Many of the people included in the DWP data could be short term migrants who do not intend to stay in the UK for 12 months. However, it could also be the case that part of the difference could be explained by the IPS underestimating the number of Polish citizens immigrating to the UK. There is no way of taking this into account within the log-linear framework detailed in this chapter.

The migration of Indian citizens to the UK is long-established (chapter 2.2) and significant. According to the HESA and DWP data there also seems to be a significant number of both student and non-student migrants from India. Figure 4.7 shows the patterns in the data used and the model results, specified by equations 4.6 and 4.7 for both

the non-student and student flow of Indian citizens. Here, the graphs are placed side by side for ease of comparison, with identical scales on the Y axis.

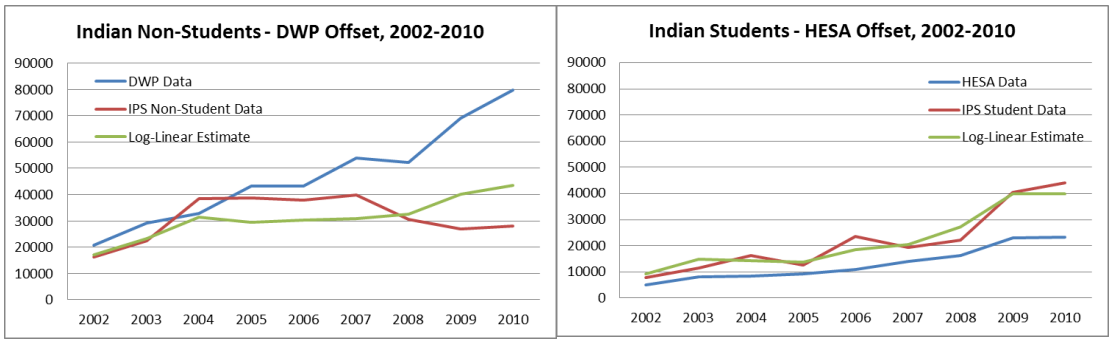

Figure 4.7: Comparison of Data Used (HESA, DWP and IPS) and Log-linear Estimates of Student and Non-Student Immigration, from Equations 4.6 and 4.7, of Indian citizens, 2002 – 2010

The student and non-student models produce similar estimates for immigration of Indian citizens. In both sources of administrative data there is a general upward trend; however, for non-students, because the log-linear estimate is constrained to the IPS margins, this trend is flattened out. The flow of Indian migrants is large, but if the pattern exhibited by the IPS estimate of non-students is driven by sampling variation, this could be distorting the log-linear estimate. Also, for non-students, there seems to be a divergence between the IPS and DWP data over time. This could be as a result of an increasing amount of shorter term migrants in the DWP data, who would not be captured by the IPS, which has a working definition based on 12 month duration of stay.

The results from the modelling seem relatively plausible when one considers the conclusions of chapter 2. Immigration from South Asian countries, such as India, is well established. Consequently, one would expect relatively stable patterns over time, as exhibited in the log-linear estimates shown in figure 4.7.

Up to this point, larger flows have been considered. The next step is to compare model results for smaller flows. Firstly, a typical European non-A8 country is considered. Figure 4.8 compares the results for Swedish citizens of the student and non-student models estimated by equation 4.6 and 4.7 and the HESA, DWP and IPS data used.

It is clear from both charts that there is a large amount of sample variation from the IPS data for students and non-students respectively; whereas, the DWP and HESA data seem to indicate that the pattern of immigration of Swedish citizens is relatively stable over time. As such, the IPS data, through consideration of the 'accuracy' assessment

criteria, is distorting true flow due to the sample noise evident in this small part of the overall IPS sample. If one assumes that the marginal totals of students and non-students for Sweden are accurate – the $\lambda_i^{IPS.S}$ and $\lambda_i^{IPS.W}$ terms in models 4.6 and 4.7 – and that the patterns exhibited by the administrative sources of data reflect the pattern of true flow, then log-linear estimates could be said to have addressed the distortion of true flow caused by the sampling variability of the IPS.

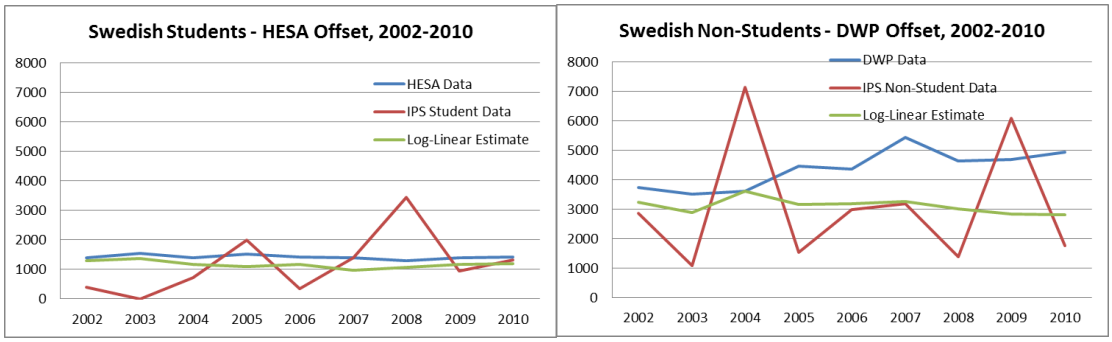

Figure 4.8 : Comparison of Data Used (HESA, DWP and IPS) and Log-linear Estimates of Student and Non-Student Immigration, from Equations 4.6 and 4.7, of Swedish citizens, 2002 – 2010

Verifying whether this is the case is problematic. Firstly, one could consider the review of main patterns of immigration from chapter 2.2. However, apart from information on the largest flows – India and China, for example – and more recent changes to immigration from the effects of EU enlargement, there is not any evidence to verify smaller flows, such as Sweden. Additionally, there is not any information that can be taken from the review of theory in section 2.3 to verify smaller flows of immigration.

A second example of a smaller flow is that of Malaysian citizens. Figure 4.9 compares the results for Malaysian citizens of the student and the non-student models and the data used.

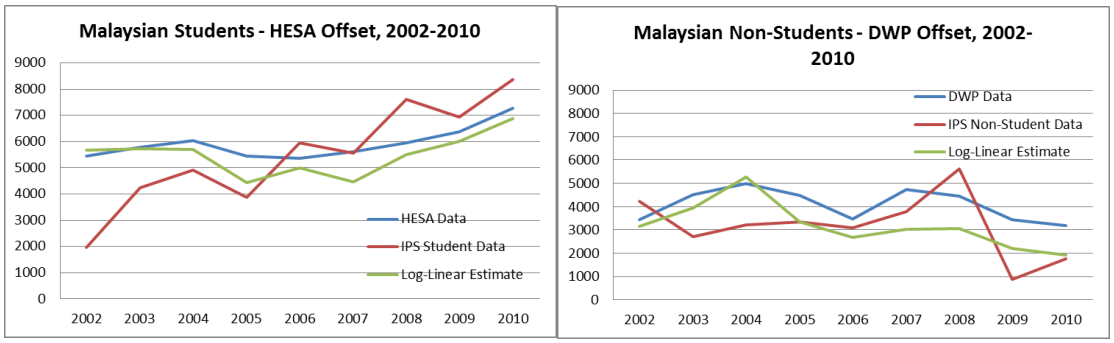

Figure 4.9: Comparison of Data Used (HESA, DWP and IPS) and Log-linear Estimates of Student and Non-Student Immigration, from Equations 4.6 and 4.7, of Malaysian citizens, 2002 – 2010

For both students and non-students, as expected, there seems to be irregularity as a result of sample error, in the IPS data. Both sources of administrative data exhibit relatively stable patterns over time. However, unlike the modelled estimated for Sweden, the fitted values for non-students estimated from equation 4.7, is more irregular. This irregularity is driven by the drop in new NINo registrations of Malaysian citizens in 2006. Again, verification as to whether this change is a good reflection of true flow is problematic due to the lack of alternative information in both previous research on quantities of immigration and the limitations of migration theory.

4.5 Conclusion

The progression of the log-linear modelling in this chapter has shown that, in order to augment the IPS estimates with two sources of administrative data that cover both students and non-student, then one has to estimate two separate models. It is also clear that log-linear models with offset terms are an effective way to combine the reliable elements of the IPS with the assumed reliable year-on-year variation of immigration taken from the DWP and HESA data sets. They are also conceptually and computationally straight forward.

For the larger migration flows – India and Poland, for example – the results from models 4.6 and 4.7 seem plausible. The introduction of an offset term smooths the IPS data to produce a log-linear estimate that reflects the pattern in the administrative sources; but, is constrained to the margins of the IPS. The marginal totals for these larger flows are deemed relatively reliable and are close reflection of true flow with regard to the definition assessment criteria. For the smaller migration flows, it would seem that the log-linear estimates smooth the very large irregularities exhibited by the IPS due to sampling error. This helps mitigate the effect of distortion of true flow caused by the sampling variation in the IPS; however, model verification is problematic with smaller flows.

Building on the suggestions of Raymer et al (2011) this approach should contribute towards the ONS fulfilling the Eurostat requirements for international migration data outlined in chapter 1. Article 3 of the European Parliament Regulation (EC) No. 862/2007 states that as part of the statistics process, scientifically based and well documented

statistical estimation methods may be used in the official submission of international migration statistics to Eurostat.

A recommendation of this chapter for the ONS is that they explore the possibility of applying the modelling approach proposed in this chapter in their estimation of citizenship specific UK immigration. Furthermore, the substantive work of this chapter formed part of a submission of evidence by Bijak et al (2013) to the House of Commons Public Administration Select Committee (HoC PASC) study of migration (HoC PASC 2013). As a result of the PASC report, based on the evidence submitted to them, the UK Statistics Authority (UKSA) have recommended to Government that alternative sources of data, such as the HESA and DWP, should be used in the estimation of UK immigration (UKSA 2013).

One of the assumptions of fitting a log-linear model to a contingency table of this kind, specifically for this research, is that the quality of the data being used does not depend on either of these two categorical variables. With regard to the IPS there is a longstanding concern that the survey design does not sample sufficiently from regional airports (ONS 2012 c). Since the expansion of freedom of movement from accession countries within the EU, it is possible that the IPS could have been undercounting migration flows from A8 countries such as Poland. This is a limitation of the log-linear framework, as there is no formal way of being able to include this in the modelling process.

A further limitation of being constrained to a log-linear framework, given the data available, is that it is not possible to describe formally in the model framework the uncertainty around the estimates of true flow. Assessments of uncertainty can only be made separately from the model results about how the accuracy of the patterns in the HESA and DWP data then being benchmarked to the IPS data has improved estimates of immigration.

It is clear that, from the conclusions above that the next part of the research needs develop a model framework which allows the data assessment of chapter 3 to be formally included in a statistical model. Furthermore this modelling framework needs to provide a coherent estimate of uncertainty. To do this it is necessary to move beyond the constraints of the log-linear models applied in this chapter. Consequently, the next chapter develops a Bayesian modelling framework that allows the inclusion of further sources of data available

– GP Registration flag 4 data and Census data – and an assessment of these sources, based on the three assessment criteria outlined in chapter 3.

Chapter 5 – Bayesian Modelling of UK Immigration Data

5.1 Introduction

In chapter 4, it became clear that there is no way of formally including subjective judgements about how specific characteristics of the collection of the available data sources have distorted the measurement of true flow in the presented frequentist version of a log-linear model. It is left to an ad hoc approach and interpretation of the results to allow for such considerations.

A further limitation of the analysis presented in chapter 4 is that the expression of uncertainty is simply a representation of the stochastic error in the model. It is not reconcilable with the level of uncertainty identified in the qualitative assessment of the data, in relation to true flow, outlined in chapter 3. Furthermore, without the survey weights for the IPS, correct statistical inference from the log-linear models is not possible. This chapter looks to address these two specific limitations, regarding modelling data assessment and uncertainty, by specifying two different Bayesian models to estimate UK immigration. Subjective judgements of the data are included, albeit in different ways, in both models explicitly and a coherent expression of uncertainty is estimated in the modelling framework.

Firstly, though, a general introduction to Bayesian statistical modelling is outlined. Through linking important elements of a Bayesian approach to this research, an explanation is provided of how a Bayesian statistical model has the potential to address the limitations discussed in chapter 4. Following this introductory section, there is a specific focus on how Bayesian models of migration have been estimated and specified previously. This review guides the specification of the two hierarchical Bayesian modelling frameworks outlined in the final part of this chapter – a Bayesian log-linear model and a data assessment model.

Key to addressing the limitation of not including an assessment of the data available explicitly in the model is the specification of prior probability distributions. It is through these prior probabilities that subjective judgements of the data sources can be modelled explicitly. As such, full probability models, for both models are outlined in detail

in this chapter. Through the specification of full probability models the framework for including the subjective judgements of the data, which are outlined in chapter 3, is established.

5.2 Bayesian Data Analysis

There are two main broad approaches to statistical inference. The first is the sampling-based frequentist approach. In theory, only events that can occur in repeatable samples are considered, with statistical inference based on the interpretation of probability related to the frequency of phenomena under study (Bijak 2010, page 27). The second, Bayesian data analysis, consists of practical methods for making inferences from data, using probability models for quantities we both observe and for quantities about which we wish to learn (Gelman et al 2004). It is based on an interpretation of probability as a rational, conditional measure of uncertainty, which closely matches how we use 'probability' in everyday language (Bernardo 2003, page 1). A fundamental characteristic of a Bayesian approach is the explicit use of probability for quantifying subjective beliefs and uncertainty in the inferences made. This is one of the key differences between the two approaches. A frequentist approach rejects the subjectivist notion of probability as a measure of belief with regard to the chances of occurrence of events, or states of nature (Bijak 2010, page 27), whereas Bayesian methods do not.

The theoretical differences between the two approaches are clear; but, there are many occasions where the two approaches are conflated. For example, in frequentist analysis, there is often a consideration of a confidence interval around a point-estimate of a parameter or predicted value. There is a long standing awareness that frequentist confidence intervals are sometimes not interpreted correctly (cf. Jaynes 1976). A confidence interval is a range of values, which frequently includes the parameter of interest if the sample is repeatedly drawn from the population (Gelman et al 2004). Whereas, a Bayesian 'credible interval' for an unknown quantity of interest – an estimate of immigration, in this case – can be regarded, in theory, as having a high probability of containing the true value of the unknown quantity.

Concentrating on the philosophical debates on the foundations of statistics is beyond the scope of this research, though. Furthermore, when one is faced with a research problem with many different types of uncertainty – about what exactly we are estimating

and the quality of the data, for example it is arguable that adherence to one particular philosophical approach could be a hindrance. A focus on idealised inference does not reflect the complex and uncertain nature of most research design, questions and data (Chatfield 2002). Chatfield (ibid), makes the case for a pragmatic approach to statistics, where considerations of the nature of the research problem are central to the methods applied.

As such, this section focuses on the advantages of applying a Bayesian framework for estimating UK immigration. Specifically its flexibility and generality make it ideal for the analysis of complex data problems, such as this research, where each source of information is collected in a different way, for different primary purposes and of varying quality with regard to estimating immigration. Furthermore, through the use of prior probability distributions to quantify our uncertainty about evidence, it is possible to include the judgements of data, set out in chapter 3. Thus, providing a coherent method to model the judgemental information in table 3.1 for the estimation of true flow of UK immigration.

5.2.1 Bayes Theorem

Statistical inference, in general, is concerned with drawing conclusions about something that is not observed, a quantity to be estimated, from data. Bayesian statistical conclusions about a parameter θ are made in terms of probability statements conditional on the observed value of data y. Using the above notation this can be written as $p(\theta|y)$ where in this case p is a probability distribtion. This is called the posterior distribution. The posterior distribution reflects the belief about the unknown value of θ given the results of the evidence gathered from the data y. Effectively, the posterior distribution is the output from a Bayesian analysis.

In order to make a probability statement about θ given y a model needs to be specified that provides a joint probability distribution for both θ and y. This is called a joint probability density function and can be written as a product of two densities - the first is referred to as the prior distribution $p(\theta)$ and the second density is that of the likelihood of the data, $p(y|\theta)$:

$$p(y, \theta) = p(\theta)\, p(y|\theta) \tag{5.1}$$

where $p(y, \theta)$ is the joint probability function. Simply conditioning on the known value of the data y using the basic property of conditional probability, Bayes' rule, yields the posterior density:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \quad (5.2)$$

Using a Bayesian approach, outlined above, statistical inference about a quantity of interest is described as the modification of uncertainty about its value in the light of evidence (Bernardo 2003) and, as such, Bayesian analysis provides researchers and statisticians with a rigorous way to make probability statements about real quantities of interest. Furthermore the prior distribution $p(\theta)$ is the opportunity to depict the knowledge and beliefs of the researcher, with respect to possible values of θ, unconditional on the empirical evidence of the data (Bijak 2010).

This additional consideration of subjective belief explicitly in a statistical model is the main motivation for applying Bayesian statistical methods to UK immigration data. If the analysis is limited to just a frequentist model, there is no way of including our additional knowledge of the characteristics of each of the sources of data, as outlined in chapter 3, explicitly in a statistical model.

5.2.2 Applied Bayesian Statistical Analysis

In an application of a Bayesian statistical model, in general, there are 3 steps (Gelman et al 2004). The first step is a *specification of a full probability model for all observable and unobservable quantities of interest.* This is better known as model specification, the main focus of this chapter, but does also include prior elicitation, the main focus of chapter 6. The second step is to *update the knowledge specified in the full probability model about the unknown parameters by conditioning on the available observed data.* This is effectively the computation of the model. This is carried out, in this thesis, using the Open Bugs software (Spiegelhalter et al 2011). The third step outlined by Gelman et al is *an evaluation of the model, which includes the sensitivity of the conclusions to the main assumptions made.*

In this section, a brief introduction to the first step in Bayesian analysis, specification of a full probability model, is outlined. The establishment of the full probability model is the most important step, and the main stumbling block, in Bayesian

analysis. Documenting and justifying where the models come from, and how we construct appropriate probability specifications for all known and unknown parameters of interest, is the essence of applying Bayesian methods.

As outlined in the previous section, broadly speaking, two probability densities need to be specified – a density for prior probabilities and a density for the likelihood of the data. For the prior probability densities this includes, if available or possible, some prior knowledge about the parameter values to be estimated; thus, enabling us to incorporate auxiliary and previous knowledge on the subject (Gill 2002). For the purpose of this research, this can take the form of alternative sources of data and judgements of how closely this data, or data used in the likelihood, matches the true immigration flow.

The likelihood is the observed data, conditional on its parameters. In practical terms, and very much relevant to this research, this can include a standard frequentist model of observed data. In frequentist analysis parameter values are viewed as being fixed, and probability statements can be applied to possible values for the data, given the parameter values. In a Bayesian analysis, however, all parameters are treated as random meaning there are probability distributions for the parameters, as well as the data (Agresti 2013). Therefore one can make probability statements about possible values for the parameters given the data. Importantly, this allows subjective judgements about each unknown parameter to be included in the model. When priors include this extra information, they are referred to as being informative. This information can come from previous studies, or from an alternative source of data or a specific judgement about the data collection process. Alternatively, the prior can be relatively uninformative, so that the results of fitting the model are based almost entirely on the data (ibid). In this case they are commonly known as 'vague' or 'vaguely informative priors'.

Prior probability densities are specified in relation to each of the parameters of the model. The prior distribution combines with the information that the data provides (likelihood) to generate posterior distributions for all the estimated parameters and fitted values, providing a coherent expression of uncertainty of all values of interest in the model. The way prior distributions are set can result in different values for the posterior distributions. Consequently, specifying the distribution of the priors and eliciting values for their hyperparameters, where a hyperparameter is a parameter of a prior distribution, should be given careful thought and thorough justification,. For example, where there is

little evidence available about the parameter of interest, then appropriate values are given to the hyperparameters to elicit a high level of uncertainty in the prior.

In the specification of the priors an appropriate probability distribution needs to be chosen; for example, whether the prior assumes a normal or a gamma distribution. For each of the two model frameworks for estimating immigration this is outlined in section 5.4. Following this one needs to quantify the relevant subjective judgements, so that the hyperparameters of the prior distributions can be set. This is known as prior elicitation and is described, explained and justified in detail in chapter 6. Elicitation is a substantial task and this is why a whole chapter is devoted to it. Bridging the gap between the data assessment in chapter 3 and the parameters specified in this chapter, upon which the priors will be set, is one of the main contributions of the prior elicitation.

The key contribution of using a Bayesian approach in this research is that it enables all publicly available evidence on UK immigration to be used in a model estimate. With this, and the challenge of the prior elicitation outlined above in mind, where possible, the models in this chapter are specified in way to make inclusion of the data assessment categories as straight-forward as possible.

In a sense the three stages of Bayesian analysis introduced at the start of this section can be thought of as a learning process. We have our perceptions of a specific issue based on our current knowledge, including the amount of certainty we are willing to attribute to this. For instance, we are relatively certain that a source of data is likely to overestimate long term immigration because short term migrants will be included in the data set. This is the case, for example, in the DWP data where short term migrants who wish to work have to obtain a NINo. However, we do not know exactly the magnitude of this over-estimation, which means an expression of uncertainty about this prior knowledge is required. This judgement, including the estimate of the overcount and the uncertainty with which the judgement is made, is the prior.

These perceptions are then updated in the light of evidence, which in this case is the data itself. So, continuing the example, we can then use the NINo data for specific countries and specific years to update our prior belief on data collection, into a posterior estimate of country-specific immigration to the UK, known as 'true flow'. Furthermore, an expression of the uncertainty of the estimate of true flow can be obtained from the

posterior distribution, which is effectively a summary of the outcome of the learning process outlined above.

However, because of the uncertainty of our prior beliefs – we do not know exactly by how much NINo data overestimates immigration, for example – we need to see how sensitive the posterior estimates are to our assumption of this distortion of true flow. Hence the third step in the application of a Bayesian model, the evaluation and sensitivity analysis to our prior assumptions.

5.2.3 Relevance of Bayesian Approach for Modelling UK Immigration

Adopting a Bayesian approach, as previously mentioned, gives the researcher the flexibility to use various sources of data in one coherent probabilistic model (Wheldon et al 2013 and Raymer et al 2013). Whereas, the use of multiple sources, for this particular research, in a standard freqentist log-linear analysis as outlined in chapter 4, is limited to ad hoc models and solutions.

Seeing as though the essence of a Bayesian approach is using *all* the evidence available to make inferences about a quantity of interest, the use of a Bayesian model to combine multiple sources of data to better estimate true flow of UK immigration addresses the two main research aims of this thesis. Rather than just relying on the IPS, the development of the Bayesian models in this chapter does not only produce estimates which make better use of all the evidence outlined in chapter 3, but also aid our understanding of the uncertainty inherent in UK immigration data.

In this study, information about the quality of the data sources that are readily available to estimate country-specific immigration flows, outlined in chapter 3, provides the basis for our prior beliefs. Through the use of prior distributions, it is clear that Bayesian inference lends itself to allowing formal considerations about data collection and definitional differences, which means we can model these judgements explicitly. An application of Bayesian models allows us to include more of the 'judgemental evidence base' of the data established in chapter 3 and in a more effective way.

The sources of UK immigration data themselves are included in the likelihood part of the model, and are what is used to modify the priors. The output of the model -

posterior probability distributions - provides a probabilistic expression of the estimates of immigration produced by the model, which includes a coherent expression of uncertainty.

A further contribution of adopting a Bayesian framework, is that through an analysis of the uncertainty estimated in the posterior distributions, one is able to further our understanding of the main sources of uncertainty (and, of course, certainty) in the available UK immigration data. In the limited previous work on UK immigration estimation, the focus has been on *improving* estimates. The motivation for this emphasis is clear and justified and has been covered in detail throughout this thesis.

However, there is a need to move beyond this approach, as in the absence of collecting better data, the statistical models applied in this and previous research can only improve estimates up to a point. One way this can be achieved is by using the analysis of the posterior distributions and the uncertainty estimated to provide assessments of the uncertainty inherent in each source of data. Furthermore, through assessing how the uncertainty in immigration estimates is affected by the uncertainty propagated from each assessment criteria, for each given source of data, recommendations about what is required from the ONS to improve their immigration estimates can be made.

It is therefore possible that a Bayesian approach could alleviate some of the difficulties faced by conventional statistical methods (Bernardo 2003), as experienced in the previous analysis, which focused on combining data in frequentist a log-linear framework. Also it could provide a contribution towards uncertainty being fully considered in the estimation of immigration and how it is used to make improvement in any future proposed changes to the design of data collection.

5.3 Review of Bayesian Models of Migration

This section briefly outlines recent applications of Bayesian estimates of migration and the main relevant contributions they make for this research. In chapter 2 there is a broad review of statistical modelling of migration data. The contribution of this section, however, is to identify relevant details in the most recent applications of Bayesian models that are useful in the model-development for this research. This review is split into two sections, mirroring the model specification in sections 5.4.1 and 5.4.2 respectively. The first section (5.3.1) outlines general considerations that need to be made in the Bayesian log-

linear analyses of contingency tables and an example where a Bayesian log-linear model has been applied to a migration data problem is outlined. The second (5.3.2) reviews measurement error models in demography, where Bayesian methods are applied to improve estimates of migration and population statistics in general. The relevant conclusions of sections 5.3.1 and 5.3.2 are taken forward into the model specification in section 5.4.

5.3.1 Bayesian Log Linear Models

The first model to be specified, to estimate UK immigration in this chapter, is an extension of the frequentist log-linear offset model, applied in chapter 4, to include priors for each of the model parameters. This model is specified in section 5.4.1. Firstly, though there is review of Bayesian log-linear models in this section.

In general, when undertaking a Bayesian log-linear analysis of contingency tables, it is necessary to specify priors for either the cell counts or the log-linear parameters (King and Brooks 2001 a). There are several disadvantages of specifying priors for the cell counts, the most obvious being that it would require a very large number of priors in comparison to specifying the priors for just the parameters in the model. Also, with regard to this research and as concluded as a result of chapters 2 and 3, the evidence available on data collection and assessment of the data sources is far stronger, and of more use, than the evidence available about what is driving the levels of country-specific immigration to the UK over time.

Furthermore, evidence available about the assessment of the data naturally lends itself to being elicited as a prior for a marginal effect of a log-linear model of a contingency table. It is unrealistic to judge the distortion of true flow for each separate cell of data, whereas providing some kind of assessment of the distortion of true flow for a given data source, for the marginal probabilities of the table is achievable (cf. Raymer et al 2013).

Therefore, for this research problem, it is more appropriate for priors to be specified for the log-linear parameters – the overall effects and main effects - rather than the cell counts. One of the advantages of this form of prior is the computational and conceptual simplicity of the model as a whole (King and Brooks 2001 a). This is because there are typically, and especially with regard to this research, far fewer parameters than

there are cells. Furthermore far few priors have to be elicited. Placing informative priors on the parameters simply expresses knowledge about the interactions and effects in the model (King and Brooks 2001 a).

Brierley et al (2008) present an exploratory Bayesian framework to estimate place to place migration flows in a migration flow table. Their framework is designed to be flexible and capable of dealing with flows of varying quality and missingness. The motivation for Brierley et al came from Raymer's (2007) paper discussed in chapter 4 where the marginal totals of a migration flow table are deemed reliable, having already been constructed. Furthermore, by adopting a Bayesian framework, they can include prior information explicitly into the modelling process.

An advantage of the framework adopted by Brierley et al, over the more ad hoc sequential approach outlined in Raymer's (2007) paper is that the model can be estimated in one direct step. This critique is also relevant to this research too; the log-linear framework specified in chapter 4 is somewhat ad hoc and two models have to be fitted separately so that offset terms can be applied to the relevant part of the IPS sample. Furthermore, inclusion of additional evidence helps address the stated aim of this research – inclusion of all useful available evidence on UK immigration. A Bayesian approach similar to the one of Brierley et al provides a suitable starting point for extending the analysis of chapter 4.

Brierley et al (2008), like Raymer (2007), are interested in estimating the origin-destination flow matrix of a migration system. As previously outlined, log-linear models of contingency tables can be used to describe the underlying structure of a contingency table, which makes their application suitable for modelling data like migration flow tables, which, it can be argued, have some kind of spatial structure. Each of the main effects in the log-linear models are interpreted as push and pull factors, respectively, for region-to-region migration. Brierley et al assume that they have a matrix of **Z** reported migration flows and a matrix **Y** of true migration flows with unknown entries, but known, reliable fixed margins (Brierley et al 2008, page 152). Initially they assume that **Z** is complete and that the relationship between **Y** and **Z** is

$$\log z_{ij} \sim N\left(\log y_{ij}, \sigma^2\right), \quad i \neq j \tag{5.3}$$

where $\log z_{ij}$ are independent with a common variance and where $i \neq j$ indicates that the model takes into account the structural zeroes on the diagonal of the migration flow table.

A log-normal distribution is specified; however, a Poisson distribution model for the observed counts could be considered more appropriate (ibid, page 153). An assumption made here, and one that will have to be made in any Bayesian extension of the log-linear model with offset specified in chapter 4, is that each of the reported values has the same accuracy. A key second assumption is that the reported values are unbiased. However, as commonly encountered in studies of migration data and illustrated from the data assessment in chapter 3, it could be the case that there is often undercount in migration data collection systems and this could be country-specific.

Matrix **Y** is the unknown migration matrix where the sums of the columns and rows are assumed known, and have been estimated before fitting the model. The true migration flows, the values to be estimated, $\log y_{ij}$, follow a prior distribution which is centred on a log-linear model assuming quasi independence:

$$\log y_{ij} \sim N\left(\mu + \alpha_i + \beta_j, \sigma^{*2}\right), \quad i \neq j \tag{5.4}$$

where μ is the overall effect, α_i is a pushing effect, β_j is a pulling effect and σ^{*2} represents the variance in the prior probability distribution which is centred on the quasi-independence model. Initially for each parameter in the quasi-independence model Brierley et al (2008) assume normal prior distributions. King and Brooks (2001 b) also place independent normal priors on the log-linear parameters in their contingency table analysis of population size. Brierley et al (2008) then go on to specify inverse gamma distributions for the σ^2 and σ^{*2} hyperparameters respectively. The priors are conditionally conjugate with likelihood, which makes computation straightforward.

All of the prior distributions are initially set to be diffuse, which means that the parameters from the independence model are well supported (ibid, page 154). This results in posterior estimates which are driven by the log-linear model quasi-independence specified by equation 5.4. Through estimating the model with vague priors, the researcher gains a posterior estimate of what we learn from the data, given the likelihood model specified.

When informative priors are specified, and the model is re-estimated, one can compare the posteriors computed from vague and then informative priors. The difference in the posterior estimates here provides an estimate of how sensitive the data is to prior judgements made. In the case of Brierley et al (2008), they propose including informative

priors which capture the judgement of the quality of the reported data (ibid, page 161). As such, informative priors are specified for σ^2 and σ^{*2}. These hyperparameters capture the variance of the distributions represented by equations 5.3 and 5.3, for the quasi-independence model and the reported data respectively.

So, Brierley et al argue that through specifying informative priors, one can express prior beliefs on the accuracy of how well migration data could be expected to conform to the fit of the quasi-independent model or experts could give prior beliefs regarding the accuracy of the reported values. Brierley et al do not, however, specify informative priors for the main effects parameters in the quasi-independence model, as the margins for this model are assumed reliable.

The main relevant conclusions that can be taken from the proposed Bayesian approach outlined above, is that a Poisson distribution for the likelihood is an appropriate specification for that part of the model. Another element of their research which can be taken forward into the model specification is the use of independent normal priors. However, the Bayesian log-linear model of UK immigration specified in section 5.4.1 differs slightly in that informative priors are specified on the parameters in the log-linear model of independence, whereas vague priors are specified on the equivalent parameters by Brierley et al.

A similar assumption to Brierley et al is made in this chapter, in that the IPS margins are deemed relatively reliable, but they are not fixed. The informative priors on the main effects in the model aim to assess how sensitive the log-linear framework is to the inclusion of further auxiliary data. Finally, the log-linear parameters cannot be interpreted in the same way as Brierley et al did. Firstly, because the latter are modelling a migration flow table which has a spatial structure, which the citizenship-by-time contingency table of immigration does not have. The inclusion of the offset term further complicates parameter interpretation. Consequently, even though the Bayesian log-linear model of UK immigration is relatively similar to the Brierley et al's (2008) model outlined by equations 5.3 and 5.4, there is still work to be done on parameter interpretation to help set the priors and a model specification which allows informative priors to be set on the main effects of independence model.

5.3.2 Measurement Error Models

In recent years there have been few examples of Bayesian modelling of international migration data, which take into account differences in availability, quality and definitions. Most of this research has looked at modelling a multi-country problem (cf Raymer et al 2013). The analysis in this chapter, though, attempts to apply a modelling approach that has so far just been applied only to multiple country flows – modelling of an international migration matrix – to a single country research problem: international migration to the UK. The same principles applied to a multi-country problem, however, can be used in the specification of the data assessment model in section 5.4.2. Where differences with regard to definition and coverage, for example, are taken into account for the official source of data on migrants sent and received, for each country, in a migration flow table by Raymer et al (2013), differences in multiple sources of data can be taken into account in a similar way in the single country problem of this research.

Raymer et al (2013) propose a Bayesian model for migration between European countries, which attempts to both harmonise and correct inadequacies in available data and to estimate completely missing flows. The motivation for their research is that a Bayesian approach will allow the opportunity to integrate different types and multiple sources of data, migration theory and expert judgement about data collection into one coherent method of estimation (ibid). Their approach is reviewed in detail below and some of the characteristics of their model specification are used in the data assessment model outlined in section 5.4.2.

The focus of Raymer et al is on estimating a migration flow table of unobserved true flows based on four pieces of information: flows reported by the sending country, flows reported by the receiving country, covariate information and expert judgements (Raymer et al 2013, page 802). Their model can be thought of in two parts – a migration model based on theory and a measurement model. The theory model is used to augment the measurement and to estimate any missing flow data. As determined in chapter 2, the research in this thesis is data driven, with theory only used to verify some results in an ad hoc way. As result, the focus of this review is on the measurement error part of the model used by Raymer et al.

The reported data are harmonised, so that they reflect the concept of true flow, via two measurement error equations. Firstly though, Raymer et al specify the distribution for

the data, the observed flows for both the sending and receiving countries are assumed take a Poisson distribution:

$$z_{ijt}^{S} \sim P\left(\mu_{ijt}^{S}\right), \qquad z_{ijt}^{R} \sim P\left(\mu_{ijt}^{R}\right). \tag{5.5}$$

Sub-scripts *i* and *j* denote the origin and destination in the migration flow table and *t* denotes the year that the flow is reported for. Superscripts S and R denote whether the data is reported by the sending country or the receiving country.

Raymer et al (2013, page 803) then specify two measurement error equations, where they convert the reported data to comply with the UN definition:

$$\log \mu_{ijt}^{S} = y_{ijt} + \delta_{m(i)} - \log \lambda_{f(i)} - \log(1 + e^{-\kappa_i}) + \varepsilon_{ijt}^{S} ,$$

$$\log \mu_{ijt}^{R} = y_{ijt} + \delta_{m(j)} - \log \lambda_{g(j)} - \log(1 + e^{-\kappa_j}) + \varepsilon_{ijt}^{R}. \tag{5.6}$$

The term y_{ijt} denotes the true flow of migration from country *i* to country *j* in year *t*. The true flow term here is the same as the one detailed in chapter 3 and used throughout this thesis; it is based on the UN definition of long term international migration. It is also common term for both the sending and receiving equations, thus providing one migration flow estimate of true flow based on both sending and receiving countries data. This approach is applied in the data assessment model in section 5.4.2 to make use of multiple sources of data to produce one table of estimates.

Differences in duration of stay are taken into account by the $\delta_{m(i)}$ parameter and the effects of undercount are captured by $\lambda_{f(i)}$ and $\lambda_{g(j)}$. The duration of stay parameter takes into account the effect of there being no time limit, 3 months, 6 months, 12 months and permanent duration of stay for migrants in particular country-specific sources of data. For the undercount parameters, according to the value of $f(i)$ undercount of emigration is either assumed low or high and for $g(j)$ immigration is assumed either low or high. The classification of countries into either high or low undercount for immigration and emigration is determined by expert judgement and various studies into data collection systems in Europe (see Kupiszewska and Wiśniowski 2009, cited in Raymer et al 2013). The κ_i parameters are normally distributed random effects, which have country-group specific means and variances. This takes into account the level of coverage for given country's data source. The five Nordic countries and the Netherlands, are assumed to have excellent coverage, the mean for the distribution of the random effect is constrained to be

1 on the natural scale, which ensures identifiability for the remaining countries which are assumed to have standard coverage (Raymer et al, 2013). The error terms in the model, denoted by ε_{ijt}^{S} and ε_{ijt}^{R} are assumed to take a normal distribution centred on 0 with a variance that captures the overall accuracy of a country's data collection system. Various data collection systems are split into three categories – excellent Nordic registration systems, other good registration systems and poor registers or survey data.

Each of the parameters is specified in a way that either deflates or inflates the true flow parameter appropriately. For example, if the source of data is deemed to undercount migration then the parameter needs to have the opposite effect, as the true flow term is on the same side of the equation. Effectively, true flow is a balancing term in the equation which expresses all of the distortion of the various characteristics of data collection taken into account by the parameters in the models specified by equation 5.6. This principle will be used in the specification of the data assessment model in section 5.4.2. The general approach of the Raymer et al measurement error equations – capturing the distortion of true flow through the specification of appropriate parameters upon for which prior probabilities can be elicited - is taken forward to the specification of the data assessment model in section 5.4.2.

5.4 Specifying Models for UK Immigration Data

In this section two Bayesian models of UK immigration are specified – namely a 'Bayesian Log-linear Model' and a 'Data Assessment Model'. Firstly though, the motivation for specifying these models is outlined below in brief.

In chapter 4 a log-linear model was used to combine administrative sources of data, with the IPS, to estimate true flow. As previously stated, though, this model is limited in that it does not allow judgements of data to be modelled explicitly and the uncertainty, because of the limitations of the data available, was simply a reflection of the stochastic error in the model. Throughout this chapter it is clear that a Bayesian approach, through its coherent treatment of uncertainty as probability distributions on all quantities of interest (known and unknown) and the potential for prior information to include subjective judgements of the data, can address these limitations.

Building on the analysis of chapter 4, a Bayesian log-linear model is specified, where the model applied in chapter 4 is specified here as a full Bayesian log-linear model. This has the potential to allow the inclusion of sources of data that have remained unused thus far, as priors on the log-linear parameters. Including these unused auxiliary sources of data in a Bayesian log-linear framework allows us to test how sensitive the data is to the inclusion of auxiliary information on the marginal effects. This model is specified in the following section, 5.4.1. The main relevant conclusions from section 5.3.1 help guide the model specification.

The model specified in section 5.4.1, however, is still constrained to the log-linear framework; albeit augmented by some subjective judgement of extra data sources as marginal effects. As such, and following Raymer at al (2013) an alternative model, namely the data assessment model, is specified in section 5.4.2. This model is explicitly and directly based on the assessment criteria taken from chapter 3 and is closer to a model that purely represents the assessment of the available sources of data, as summarised in table 3.1.

The parameters of the model are specified to capture the distortion of true flow that can be attributed to each of the data assessment criteria – definition, coverage, bias and accuracy. As such, it is possible to gain a greater understanding of the uncertainty propagated from characteristics of data assessment for each of the sources of data.

5.4.1 Bayesian Log-linear Model of UK Immigration

The framework outlined in this section uses a Bayesian log-linear model to estimate the true flow of immigration to UK for the top 10 flows (plus an all-other category), over the period 2002-2010, separately for students and non-students.

Following the discussion of applying Bayesian methods in section 5.2.2 the first step is to specify a full probability model for all known and unknown quantities of interest. There are two stages to this. The first is assuming a distribution for the data and specifying a log-linear independence model of that data; this makes up the likelihood part of the full probability model. Secondly the prior distributions for each of the parameters of the log-linear model need to be specified.

The data takes the form of two matrices of reported migration flows, for both the student and non-student models and is taken from the IPS. The IPS is split up into students and non-students in the same way that it is in chapter 4. However, it is only the top ten flows (plus an all-other category) that are to be estimated. It is assumed that the reported data, for students and non-students follow Poisson distributions:

$$y_{it}^{S} \sim P\ (\mu_{it}^{S}) \tag{5.7}$$

$$y_{it}^{W} \sim P\ (\mu_{it}^{W}) \tag{5.8}$$

with means μ_{it}^{S} and μ_{it}^{W} and where y_{it}^{S} and y_{it}^{W} are independent observations. Subscript *i* denotes the country of citizenship and subscript *t* denotes the year of the flow; and, superscripts *S* and *W* denote student and non-student respectively. The top ten flows are determined in the same way as the top 30 flows are selected in chapter 4 – it is the ten citizenships with the largest total immigration over the whole study period, 2002-2010, according to the IPS. These countries are (in alphabetical order) Australia, China (excluding Taiwan), France, Germany, India, Pakistan, Philippines, Poland, South Africa, USA and the All Other category.

A Poisson distribution is chosen here because its mean is strictly positive - immigration counts have to be non-negative - and it also reflects the probability of immigration counts occurring each year if we assume a known rate of immigration and independence since the last event. Another assumption being made is that each of the observed values for immigration from the IPS has the same level of accuracy and that it is unbiased. As mentioned in section 5.3.1 this is not always the case with international migration data. From chapter 3, we also know that there could be under-coverage at regional airports, which is an issue in estimating migration from accession countries following the expansion of freedom of movement within the EU (ONS 2012). This limitation is taken into account in the interpretation of the results.

The next stage of specifying the full probability model is to outline the parameters for the Bayesian log-linear independence model. The two models for students and non-students, respectively, are below:

$$log\left(\mu_{it}^{S}\right) = \lambda^{S} + \lambda_{i}^{S} + \lambda_{t}^{S} + log\left(n_{it}^{HESA}\right) \tag{5.9}$$

$$log(\mu_{it}^{W}) = \lambda^{W} + \lambda_{i}^{W} + \lambda_{t}^{W} + log(n_{it}^{DWP}). \tag{5.10}$$

These two models are effectively the same as the frequentist log-linear models specified by equations 4.6 and 4.7 in chapter 4. Where $\lambda^S$and $\lambda^W$are overall effects, λ_i^S and λ_i^W are country-specific row effects and λ_t^S and λ_t^W are year-specific column effects. Both models are constrained to the more reliable marginal totals in the IPS and, as in chapter 4, the model uses the corner-cell constraint. How this affects interpretation of the parameters is outlined in detail in the following chapter.

The model splits the IPS data into two flows - non-students and students – to allow the administrative data from the HESA data of student records and NINo registration DWP data to map onto the IPS data. As in chapter 4, the offset terms $log(n_{it}^{HESA})$ and $log(n_{it}^{DWP})$ use data from the HESA and DWP administrative sources, respectively.

Using a frequentist log-linear model, with administrative data available for the offset terms means that two separate models had to be estimated. The values to be estimated, true flow, is assumed to take a Poisson distribution:

$$\hat{y}_{it}^S \sim P(\mu_{it}^S) \text{ and } \hat{y}_{it}^W \sim P(\mu_{it}^W), \tag{5.11}$$

where $\hat{y}_{it}^S$ and $\hat{y}_{it}^W$ denotes estimates of student and non-student true flow and μ_{it}^S and μ_{it}^W are the respective means of the student and non-student models of independence, specified by equations 5.9 and 5.10.

However, one of the advantages of taking a Bayesian approach is that the predicted values of models specified by equations 5.9 and 5.10 can be simply added together for students and non-students to get an overall true flow for both student and non-student immigration:

$$\hat{z}_{it} = \hat{y}_{it}^S + \hat{y}_{it}^W, \tag{5.12}$$

where $\hat{z}_{it}$ is the estimated, overall true flow. The posterior distribution of this simple addition provides an estimate of the uncertainty for total UK immigration, which includes all of the uncertainty propagated from both likelihood models outlined above and the priors outlined below.

Referring to figure 5.1, it is clear that the data is generated by a Poission process with mean μ_{it}^S and μ_{it}^W. In turn, these means are a function of the likelihood part of the model, which is specified by equations 5.9 and 5.10, and is detailed by the respective λ

parameters on the extreme left and right of the diagram. The offset terms are fixed with zero variance and perform the same function as in the chapter 4.

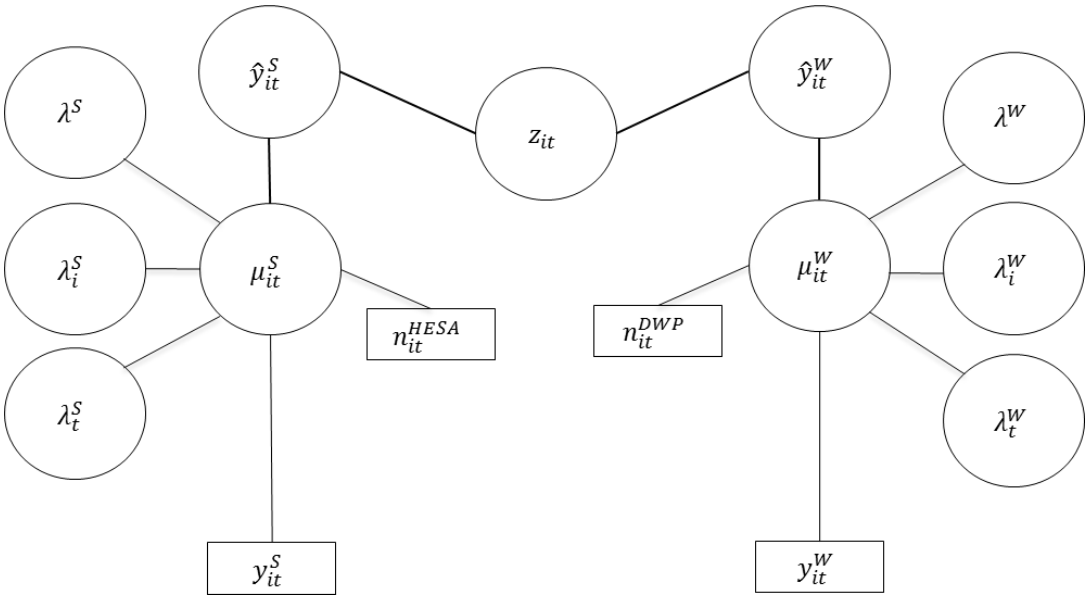

Figure 5.1 Graphical representation of the Bayesian log-linear model of UK immigration. The hyperparameters of specified for each of the prior distributions are not shown for clarity. Subscripts *i* and *t* denote citizenship and year respectively. Superscripts *S* and *W* denote whether it is the student or non-student model respectively. Square boxes denote the data and offset terms and the circles denote parameters in the model.

Moving on to specification of the priors, it is firstly important to consider their role in the joint probability model. Each prior has to relate to the parameters specified in models 5.10 and 5.11. The expected cell counts of the likelihood models link to the explanatory terms in the model – the overall effects and main effects - using the log-link. Consequently, as the parameters of a log-linear model are additive on the log scale, the priors also need to be specified on the log-scale. As mentioned in section 5.3.1 in similar previous studies independent normal priors have been specified on the parameters of a log-linear likelihood. As such, normal distributions are specified for the overall and main effects for the models set out by equations 5.9 and 5.10.

Throughout the thesis, a normal distribution is specified as $N(\theta, \tau)$,with mean θ and precision τ, where τ is equivalent to inverse variance, or $\frac{1}{\pi^2}$, where here π^2 is the variance. This parameterisation of the priors matches that of the software chosen to fit the models, OpenBugs (Spiegelhalter et al 2011). Advantages of using this form of prior are the computational and conceptual simplicity when applied using a log-linear model.

Each of the main effects priors can then be used to include extra subjective information and further auxiliary data, in addition to the data already included in the likelihood models 5.3 and 5.4. The elicitation of the subjective information is carried out in chapter 6. However, as previously mentioned, it is important to bear in mind that each of the priors is set in relation to a parameter in a log-linear model.

The main effects priors only relate to the marginal probabilities of the citizenship-year contingency table. Therefore, independent auxiliary information about the marginal totals for the ten countries chosen and the total flow for each year is required. With this in mind, this is the reason that the Bayesian log-linear model is limited to the top ten flows as country-specific assessments of data are required in a log-linear framework. It is unrealistic for priors to be elicited, for large numbers of individual flows, which allow assessment of and inclusion of auxiliary data for each separate country.

All the main effects priors in both likelihood models will be centred using the following auxiliary sources of data. For priors of the country specific main effects, λ_i^S and λ_i^W, the 2001 Census is the only source of data, which is independent from the data used in the likelihood models, available. From chapter 3, the limitations of this source of data are clear – it is only an estimate of immigration for 2001, when the study period is 2002-2010; and, it is transition data rather than the flow data of the IPS. There are also potential biases in the census data, in that there are groups in society who are hard to count and might not complete the census form. Finally, there is no way to split the Census data into students and non-students, with the same data having to be used to inform the prior elicitation for both likelihood models.

For priors of the year specific main effects, λ_t^S and λ_t^W, NHS Flag 4 data is the only source of data, independent of the data used in the likelihood models, available. In short, this data set is the number of NHS GP registrations of foreign nationals in England and Wales. The first limitation of this source of data is that it lacks the coverage of migrants who register with a GP in Scotland. Furthermore, it is reliant on self-registration. Certain migrant groups are less likely to register with their GP (Raymer et al 2011 a). Similar to the 2001 Census data used for the country specific main effect priors, there is no way separating the Flag 4 GP registration data into students and non- students.

The main limitations of these data sources with regard to how effectively they estimate the marginal true flows are clear and have been outlined in brief. They will be

considered in more detail during the prior elicitation in chapter 6 and these limitations will be reflected in the uncertainty expressed in the prior distributions.

This model, through the use of informative prior probability distributions, does allow the inclusion of further sources of data and an explicit consideration of some of the uncertainty present in the auxiliary sources of data in estimating true flow. However, ultimately it is constrained to the log-linear framework of chapter 4. The posterior estimates are still, in a way, constrained to the IPS margins. If they are sensitive to the priors specified, then the relative balance of the marginal totals could change, so they are not fully constrained to the IPS margins. However, through using a log-linear framework, overall, the posterior means are still constrained to IPS global totals. It is just the balance between country specific and year specific flows that will change if the results are sensitive to the selection of the prior distributions.

Furthermore, the Bayesian log-linear model also does not make the best use of the HESA and DWP data and the judgements of these sources outlined in chapter 3, in the estimation of true flow. Through confining the HESA and DWP data to the offset terms, we only use this administrative data to impose the year on year variation on citizenship specific immigration estimates. These two sources of administrative data are simply used in the estimation of the patterns of true rather than fully utilised to estimate the quantity of migration from each citizenship of interest, over time.

Finally, the overall and main effects parameters specified are not direct assessments of the distortion caused by data collection to true flow. Subjective judgment about the distortion of true flow, via the use of alternative marginal probabilities obtained from the census and Flag 4 registration data, is constrained to the log-linear model. To make better use of the HESA and DWP data, and to move beyond the log-linear constraints of this model, it is clear that a model which is unconstrained and fully utilises all the evidence available, both data and the information outlined in the data assessment of chapter 3 in the estimation of each cell in the contingency table, is required.

5.4.2 Data Assessment Model

The Bayesian data assessment model aims to address the broad limitations outlined above of the Bayesian log-linear model. It moves beyond the constraints of the log-linear

framework and models the assessment of each source of data chosen in relation to true flow. To achieve this, a fully Bayesian hierarchical model is specified where parameters used to estimated true flow match the assessment criteria outlined in chapter 3. As a result, the data assessment model, specified in this section, models the distortions of true flow, explicitly, as summarised in table 3.1 in chapter 3.

The Bayesian model outlined in this section estimates the true flow of immigration to UK for the top 30 flows (plus an all-other category), over the period 2002-2010, for students and non-students. The model uses the three most comprehensive sources available – IPS data separated into students and non-students, HESA data on non-UK domiciled students and the DWP NINo registration data of foreign nationals.

As with the Bayesian log-linear model the purpose of model specification is to state a full probability model for all known and unknown quantities of interest. A graphical representation of the model is set out in figure 5.2. There are three main layers of hierarchy to consider. The first layer is where the data, y, enters the model and the consequent first step in the model specification is assuming a probability distribution for the data; this makes up the likelihood part of the full probability model. The data is represented by the rectangular shapes in figure 5.2.

The second layer of hierarchy is to specify parameters which capture the distortion of true flow as summarised in table 3.1. A true flow parameter is also included in this layer of the model, denoted by z. These parameters explicitly match the data assessment criteria outlined in detail in chapter 3 and are referred to as data assessment equations throughout this section. They provide a layer in the model where prior distributions can be specified.

The priors are the third layer of hierarchy. Through assuming prior distributions for each of the parameters on can quantifying the distortions of true flow and also allowing expert judgements of the four sources of data to be explicitly included in the model. This is carried out by eliciting values for the hyperparameters of the prior distributions in chapter 6. For clarity, these hyperparameters are not included in figure 5.2.

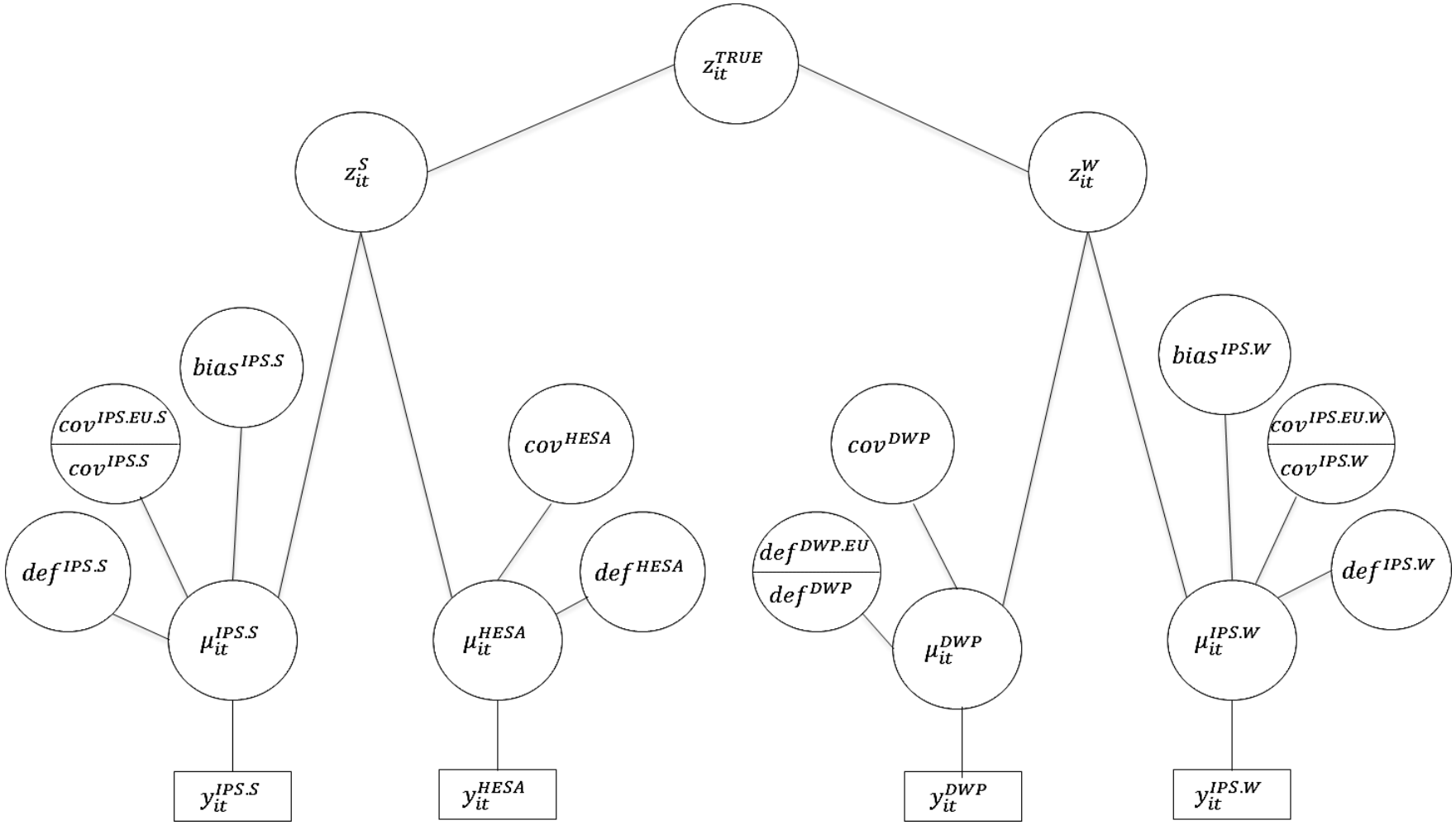

Figure 5.2. Graphical representation of the Data Assessment Model. The hyperparameters specified for each of the prior distributions are not shown for clarity. Subscripts *i* and *t* denote citizenship and year respectively. Superscripts denote the source of data each parameter refers to. Square boxes denote where data is introduced to the model. Parameters are split into two where there is an EU/Non-EU split in that particular assessment of the data. For graphical clarity the μ terms are not split into EU/Non EU; this is explained further in the model specification below. The z terms in the model are the true flow parameters.

The neat conceptual separation of the likelihood and priors is somewhat blurred in this model, a common occurrence in Bayesian modelling. Indeed, often in a hierarchical model, it is unclear at which point the likelihood ends and the prior begins (Garthwaite et al 2005). So, for the purposes of the remainder of this section, the likelihood refers to the probability distribution assumed for the sources of data. The term 'data assessment equations' refer to the parameters in the model that allow the data assessment criteria to be explicitly modelled; and the prior distributions are predominantly where expert judgement is included in the model, through their respective hyperparameters.

The data take the form of four 31 x 9 matrices of reported migration flows. It is assumed that the reported data, y, in each of the four matrices follow Poisson distributions:

$$y_{it}^{IPS.S} \sim P(\mu_{it}^{IPS.S}) \qquad y_{it}^{HESA} \sim P(\mu_{it}^{HESA})$$

$$y_{it}^{IPS.W} \sim P(\mu_{it}^{IPS.W}), \qquad y_{it}^{DWP} \sim P(\mu_{it}^{DWP}) \tag{5.12}$$

with mean μ and where y denotes independent observations of the reported data. The subscript *i* denotes citizenship, subscript *t* denotes the year and the superscripts denote the source of data that is being referred to – *IPS.S* is the student part of the IPS and *IPS.W* is the non-student part of the IPS, with *HESA* and *DWP* being self-explanatory. The top 30 flows are the same as chapter 4, chosen because they have the largest flows, according to the IPS, over the whole study period of 2002 – 2010, plus an All-Other category. Poisson distributions are chosen for the same reasons given for their use in the Bayesian log-linear model in section 5.4.1.

The next step is to outline the data assessment equations for each source of data, which allow the quantification of the distortion of each data set, in relation to true flow, to be explicitly included in the model. Parameters are specifically chosen which match the data assessment criteria outlined in chapter 3 and referred to throughout the thesis. To aid specification, the main messages of the data assessment in chapter 3 need to be considered. How, in general terms, true flow is distorted for each source of data and for each parameter guides the specification of the data assessment equations.

The data assessment equations are linked to the data in the following way. Following Raymer et al (2013), the respective hyperparameters for each of the data Poisson distributions, specified by equation 5.12, are themselves assumed to take a log-normal

distribution. These log-normal distributions are then centred on the data assessment equations and as such it is necessary that each of these parameters is specified on the log-scale. These linkages are represented in a graphical format by figure 5.2.

Included in the data assessment equations are two true flow terms; one for students and one for non-students. Similar to Raymer et al (2013), the student and non-student true flow terms act as a balancing term to the data assessment parameters in the equations. The posterior distributions of the true flow parameters provide estimates of immigration that take into account all of the distortion and uncertainty as result of the data assessment criteria from the available sources of data. This is described in more detail below.

5.4.2.1 Student Data Assessment Equations

Firstly, the student sources of data assessment equations and their priors are specified. The distortion of true flow by the data collection process for student IPS data is captured by equations 5.13 and 5.14. Seeing as though the IPS data is collected with the UN definition in mind, and referring back to table 3.1 in chapter 3, the definition parameter for IPS-student data is deemed to be a good reflection of true flow. As such the parameter does not need to inflate or deflate the data and a normally distributed prior, $def^{IPS.S} \sim N(0, \tau_{def}^{IPS.S})$, can be specified, which is centred on zero. Note carefully that for the data assessment priors throughout this section the subscripts used for the hyperparameters denotes the data assessment criteria and the superscripts denote the source of data that the prior is specified for.

With regard to survey coverage, it is necessary to group the citizenships into two groups, EU and non-EU respectively. There are 13 EU countries out of the top 30 flows, so it is approximately an even split. As outlined in chapter 3, the IPS has been deemed to be under-sampling at regional airports, leading to under-coverage of EU citizens, as EU migrants are more likely to use smaller airports rather than the hub airports such as Heathrow (ONS 2012). Consequently, for non-EU students the following data assessment equation is specified:

$$\mu_{it}^{IPS.S} \sim lnN(log(z_{it}^{S}) + def^{IPS.S} + cov^{IPS.S} + bias^{IPS.S}, \tau^{*IPS.S}) \quad (5.13)$$

and for EU students the equation is specified as follows:

$$\mu_{it}^{IPS.EU.S} \sim lnN(log(z_{it}^{S}) + def^{IPS.S} - cov^{IPS.EU.S} + bias^{IPS.S}, \tau^{*IPS.S}). \quad (5.14)$$

The use of the superscript $IPS.EU.S$ indicates that the parameter is applies to EU citizenships, for the IPS Student data. There is a minus sign in front of the $cov^{IPS.EU.S}$ parameter as there is undercount in the IPS data. The priors for both the coverage parameters are normally distributed, $cov^{IPS.S} \sim N(\theta_{cov}^{IPS.S}, \tau_{cov}^{IPS.S})$ and $cov^{IPS.EU.S} \sim N(\theta_{cov}^{IPS.EU.S}, \tau_{cov}^{IPS.EU.S})$. The values for the prior mean and precision hyperparameters are elicited from available evidence in chapter 6.

Note that the definition, bias and true flow terms are not split into EU and non-EU country groups; there is a common parameter for each of the top 30 (plus 'all-other') citizenships. Furthermore, the true flow parameter has subscripts *i* and *t* which denote that an estimate of true flow is made for each cell in the 31 x 9 contingency table. This is the case for each of the data assessment equations.

The bias term in the equation takes into account any distortion caused by migrant and visitor switchers, as the IPS an intentions based survey (see chapter 3 for full discussion). Again, normally distributed prior is specified for this parameter, $bias^{IPS.S} \sim N(\theta_{bias}^{IPS.S}, \tau_{bias}^{IPS.S})$.

A prior with a gamma distribution is set on the precision hyperparameter, $\tau^{*IPS.S}$. To avoid confusion in the prior specification, the superscript $*IPS.S$ includes an asterix to make it clear that we are referring to the overall precision term of the data assessment equations. This prior captures the accuracy assessment criteria of the IPS data and is assumed to be vaguely informative to begin with, $\tau^{*IPS.S} \sim \Gamma(0.1, 0.1)$. Throughout, the thesis gamma distributions are specified as $\Gamma(r, \alpha)$, where r is the shape parameter α the scale parameter so that the mean is $\frac{r}{\alpha}$ and the variance is $\frac{r}{\alpha^2}$. This follows the parameterisation specified in OpenBugs (Spiegelhalter et al 2011). A full discussion of these hyperparameters can be found in the prior elicitation in chapter 6.

For the HESA data, the following data assessment equation is specified:

$$\mu_{it}^{HESA} \sim lnN(lo\,g(z_{it}^{S}) + def^{HESA} - cov^{HESA}), \tau^{*HESA}). \quad (5.15)$$

The student true flow term, z_{it}^{S}, here is the same parameter here as it is in equations 5.13 and 5.14. This is illustrated by figure 5.2 where z_{it}^{S} is linked to both $\mu_{it}^{IPS.S}$ and μ_{it}^{HESA}. This is an approach adopted from Raymer et al (2013). Effectively, this produces an estimate of student immigration which is based on both sources of data and all the assessment of both sources of data, and the associated uncertainty, from equations 5.13, 5.14 and 5.15.

The definition parameter in equation 5.15 needs to deflate the data towards true flow, as we know that there are students included in the HESA data set who drop out of education within one year of commencing their studies. As such, a prior with a gamma distribution is specified, $def^{HESA} \sim \Gamma(r_{def}^{HESA}, \alpha_{def}^{HESA})$. An explanation for how the values for both hyperparameters are elicited is included in chapter 6.

As with the IPS-student model, the coverage parameter needs to inflate true flow. This is because the HESA data only includes students who are registered at public Higher Education institutions. A normally distributed prior, $cov^{HESA} \sim N(\theta_{cov}^{HESA}, \tau_{cov}^{HESA})$ is assumed for this parameter. Finally, the precision term for equation 5.15 is assumed to take a vaguely informative gamma prior, $\tau^{*HESA} \sim \Gamma(0.1, 0.1)$.

5.4.2.2 Non-Student Data Assessment Equations

Secondly the non-student data assessment equations and their priors are specified. As with the IPS student data the non-student IPS data needs to be split into two groups – an EU citizenship group and a non-EU group. The data assessment equation for non-EU citizenships is as follows:

$$\mu_{it}^{IPS.W} \sim lnN(log(z_{it}^{W}) + def^{IPS.W} + cov^{IPS.W} + bias^{IPS.W}, \tau^{*IPS.W}), \quad (5.16)$$

and for EU citizenships:

$$\mu_{it}^{IPS.EU.W} \sim lnN(log(z_{it}^{W}) + def^{IPS.W} - cov^{IPS.EU.W} + bias^{IPS.W}, \tau^{*IPS.W}). \quad (5.17)$$

As with the student data the prior for the definition parameter is assumed to be normally distributed and is centred on zero, $def^{IPS.W} \sim (0, \tau_{def}^{IPS.W})$.

Furthermore the coverage parameter, for EU citizenships, needs to take into account the undercount in the IPS inflate true flow parameter z_{it}^{W} for the same reasons

outlined for the student coverage parameters. The superscript $IPS.EU.W$ denotes the EU-specific coverage parameter. As such, there is a minus sign in front of the parameter. The priors for both coverage parameters are assumed to be normally distributed, $cov^{IPS.W} \sim N(\theta_{cov}^{IPS.W}, \tau_{cov}^{IPS.W})$ and $cov^{IPS.EU.W} \sim N(\theta_{cov}^{IPS.EU.S}, \tau_{cov}^{IPS.EU.S})$.

The bias parameter performs the same role as for the student IPS data assessment equations and the priors are assumed to be normally distributed too, $bias^{IPS.W} \sim N(\theta_{bias}^{IPS.W}, \tau_{bias}^{IPS.W})$. Finally, the prior for the precision term, which captures the overall accuracy for the non-student IPS data, takes a vaguely informative gamma distribution, $\tau^{*IPS.W} \sim \Gamma(0.1, 0.1)$.

For the DWP data it is also necessary to specify two data assessment equations, as there is an EU-specific effect from the definition assessment criteria. For non-EU citizenships the equation takes the form:

$$\mu_{it}^{DWP} \sim lnN(log(z_{it}^{W}) + def^{DWP} - cov^{DWP}), \tau^{*DWP}) \quad (5.18)$$

and for EU citizenships:

$$\mu_{it}^{DWP.EU} \sim lnN(log(z_{it}^{W}) + def^{DWP.EU} - cov^{DWP}), \tau^{*DWP}. \quad (5.19)$$

As outlined in chapter 3, there is no duration of stay criteria for people from overseas registering for a NINo. Short term migrants who intend to work need to have a NINo and are therefore included in the DWP data. Consequently, the definition parameter needs to deflate the data towards the true flow term z_{it}^{W}. An EU-specific parameter, $def^{DWP.EU}$, is specified in equation 5.19. This takes into account the effect of the expansion of freedom of movement, within the EU in 2004, from accession countries and short term migrants from the rest of the EU being included in the DWP data. Both definition parameters need to be strictly positive, as we are certain that there is an overcount in the DWP data. Therefore, a gamma distributed prior for each of the parameters is assumed, $def^{DWP} \sim \Gamma(r_{def}^{DWP}, \alpha_{def}^{DWP})$ and $def^{DWP.EU} \sim \Gamma(r_{def}^{DWP.EU}, \alpha_{def}^{DWP.EU})$ respectively.

The DWP data does not have complete coverage, as people who do not work and are not students and children do not need a NINo. However, there might be students who also work who are included, but this is deemed to be of a smaller magnitude than undercoverage outlined previously. This parameter, therefore, needs to allow of this

undercnount, so a minus sign is placed in front of it. The prior for this parameter is normally distributed, $cov^{DWP} \sim N(\theta_{cov}^{DWP}, \tau_{cov}^{DWP})$ which allows the distribution to be spread either side of zero, which, to a certain extent takes into account the aforementioned under-coverage and over-coverage.

Finally, the prior for the precision term specified in equations 5.18 and 5.19, which captures the overall accuracy for DWP data, takes a vaguely informative gamma distribution, $\tau^{*DWP} \sim \Gamma(0.1, 0.1)$.

The final part of the full probability to be specified is overall true flow term, z_{it}^{TRUE}. This is obtained from the following addition of the student and non-student true flow terms:

$$z_{it}^{TRUE} = z_{it}^{S} + z_{it}^{W}. \tag{5.20}$$

Referring back to figure 5.2, one can see that in the graphical representation of the model, the term z_{it}^{TRUE}, includes all the uncertainty propagated from all four sources of data through the likelihood and from the parameters and their respective priors from the data assessment equations.

5.5 Conclusion

An aim of large amounts of statistical research is to wring as much information from data as possible. However, in some cases, using expert opinion and all of the available evidence more effectively could add more information than the slight improvements gained through better techniques of data analysis (Garthwaite 2005). That is the main contribution of this chapter. Two modelling frameworks are specified which make better use of the available evidence, including subjective judgements, than the frequentist log-linear model of chapter 4.

Building on two specific examples of recent research, a Bayesian log-linear model and a Data Assessment Model are specified. The Bayesian log-linear model builds on the work of Brierley et al (2008) by outlining a model framework where informative priors can be specified for the marginal effects of the model of independence. Effectively the priors are an expression of quasi-alternative marginal constraints of the log-linear model.

These priors are centred using two additional sources of data – 2001 Census and Flag 4 GP Registration data. However, it is clear that just the Census and Flag 4 cannot be directly used in the priors. The auxiliary data needs to be used to set the prior means in relation to the parameters. Furthermore, an expression of uncertainty for each of the priors needs to be elicited. The level of uncertainty of these priors is a judgement of how effective the auxiliary data is as an alternative main-effects, and thus marginal constraints, in the model. Consequently, a key contribution of the following prior elicitation chapter is to bridge this gap between the interpretation of the parameters, upon which the priors are set, and the subjective judgement of the auxiliary data as alternative margins for the log-linear model.

The second model specified in this chapter looks to move beyond the constraints of the log-linear framework used up until section 5.4.1 of the thesis. A fully Bayesian Data Assessment Model is specified. This model is similar to the data harmonisation part of the model specified by Raymer et al (2013) for estimating a European migration flow table. The data assessment model specified in this chapter, however, uses data that is not necessarily designed to estimate immigration. As such, very specific distortions of true flow have had to be taken into account.

The priors and the relevant hyperparameters for the data assessment model are specified in this chapter in a way which should make their elicitation in chapter 6 less problematic than for the Bayesian log-linear model. On the other hand, the role of adequate priors is much more consequential in the data assessment model. Consequently, with regard to the data assessment model, the main contribution of the prior elicitation in chapter 6 is outlining as detailed pieces of evidence as possible that can be used to elicit prior probability distributions for the data assessment criteria.

Chapter 6: Prior Elicitation

6.1 Introduction

Throughout the thesis, in the absence of collecting more and better data to estimate the true flow of UK immigration, the case has been made for the inclusion of all available evidence on immigration, including subjective judgements of how data is collected. In chapter 5, two model frameworks are specified which allow the inclusion of subjective judgements of the data (as well the data itself), explicitly in a statistical model. This chapter explains how this subjective information is elicited into informative prior probability distributions that can be used in the respective Bayesian models outlined in chapter 5.

The main contribution of this chapter is to explain how potentially useful information, readily available, can be utilised to construct informative priors for the two Bayesian models. An informative prior expresses a specific belief (and uncertainty) about a parameter. In the case of this research, the informative priors utilise auxiliary data, evidence and expert judgement based on the data assessment criteria of chapter 3.

As mentioned in the previous chapter, with much improved computing power there has been a big increase in the scope for applying Bayesian methods to complex statistical models. These improvements also make it possible to include more realistic representations of prior beliefs. As we are less constrained by a lack of computing power, almost any prior distribution can now be combined with the likelihood (O'Hagan 1998). With this flexibility, however, comes a greater need to carefully document and justify elicitation to ensure that priors do what they are supposed to – realistically represent the subjective judgement in our case, regarding of the data, in relation to true flow of immigration, outlined in chapter 3.

In this chapter, firstly, there is a general review of how prior distributions can be elicited from subjective judgements in a way that they can be used in a statistical model. Following, this there is a discussion of prior elicitation in demographic, and more specifically, migration models. This review then guides the prior elicitation for the both the Bayesian Log-linear model and the Data Assessment Model, later in the chapter.

In general terms, this chapter acts as a link between chapters 3 and 5. In chapter 3 there is a detailed discussion of each source of available data, through the framework of the

data assessment criteria – definition, coverage, bias and accuracy. In chapter 5 two Bayesian models are specified; which, to a varying degree, allow the explicit inclusion of the data assessment of chapter 3 in the estimation process. This chapter details how we elicit these subjective judgements from chapter 3 in a suitable and appropriate way, so they can be used as priors in the models specified in chapter 5.

Specifically, in the account of prior elicitation for the Bayesian log-lienar model, a general approach to eliciting informative priors for the main effects of a Bayesian log-linear model of a contingency table is proposed. As such, a key contribution of this particular part of the research is to bridge the gap between the interpretation of the main effects parameters, upon which the priors are set, and the subjective judgement of the auxiliary data as alternative margins for the log-linear model. A simple numerical calculation is proposed which uses the 2001 Census data and the Flag 4 GP registration data in an appropriate way, given the log-linear likelihood constraints, in the elicitation of the prior medians.

In the elicitation of the uncertainty for both the log-linear and data assessment priors, a general approach, which bridges the gap between an understandable subjective judgment and the elicitation of the relevant hyperparameters of the priors is proposed. This approach uses quantiles in the tails of the prior distributions to calculate the uncertainty.

The final contribution of this chapter is the prior elicitation for the Data Assessment Model. Where possible, solid quantitative evidence is used to centre the prior distributions and to set their uncertainty; however, where this is not possible subjective judgement based on the data assessment of chapter 3 is used. Following this elicitation, and the model specification of chapter 5, posterior estimates for the Bayesian Log-Linear and the Data Assessment Model are ready to compute.

6.2 Review of Prior Elicitation

Prior elicitation is the process of formulating expert knowledge of one or more uncertain quantities as a probability distribution for those quantities (O'Hagan 2005). There are two broad aims to this – firstly to obtain sensible prior values and to secondly to provide a justifiable account of the basis for assigning these values (Goldstein 2006). It is also useful to think of the task of elicitation as a facilitation of expert knowledge into

probabilistic form. Often the expert being consulted in the elicitation is not necessarily familiar with statistical models. With the elicitation being a construction of available evidence about data collection in chapter 3, this is not a problem in this research. Nonetheless, bridging the gap between the evidence available to make subjective judgements about true flow and the parameters for which the priors are set is still a key component of the elicitation in this chapter.

In Bayesian statistics elicitation should be the basis for formulating prior probability distributions, which are in turn combined with the available data to compute a posterior distribution. However, elicitation has been more commonly used in situations where the elicited distribution will not be analysed with evidence from data, as the expert judgement encompasses effectively all the available knowledge (O'Hagan 2005). It is such applications, rather than eliciting priors for Bayesian statistical applications, which have driven the majority of research in elicitation. In section 6.2.1, there is a brief review of the main contributions for Bayesian analysis of research into elicitation, which informs the approach to elicitation for this study in sections 6.3 and 6.4.

In section 6.2.2 there is an account of how expert judgement has been elicited in various demographic models, migration estimates and forecasts. Through each of the respective reviews an appropriate general approach to prior elicitation is established, which are applied in sections 6.3 and 6.4 respectively.

6.2.1 General Approach to Elicitation

The purpose of prior elicitation in Bayesian statistics is to construct a probability distribution, that can be used in the specification of a full probability model and properly reflects the subjective judgement (and uncertainty) of the quantity of interest (O'Hagan 2006). The literature on prior elicitation, in comparison to applications of Bayesian models with informative prior information, is somewhat limited. Furthermore, it is also true that prior information can be irrelevant when the data are substantial and reliable (O'Hagan 2005). This is not the case with regard to this research, though, where the motivation for prior elicitation is clear; there are substantive deficiencies of UK immigration data. This is discussed in chapter 3, where the assessment of each of the sources of data illustrates the substantive distortions of true flow as a result of data collection characteristics. Adopting a

Bayesian approach allows us to use *all* the available evidence, and therefore our prior assessments of the data. The rationale for prior elicitation is clear.

With vastly increased computing power, over the last two decades, there has been a rapid expansion in Bayesian applications (O'Hagan 1998). The applications of Bayesian methods are vast, and social scientists can now exploit the advantages of adopting a Bayesian framework, including the explicit inclusion of subjective information in statistical models. Elicitation of expert opinion is a key task for subjectivist Bayesians (Garthwaite et al 2005). Successful elicitation of expert elicitation is a faithful representation of the opinion of the person whose expertise is being used in the statistical model (Garthwaite et al 2005).

Garthwaite et al (2005, page 5) outline four stages to prior elicitation, which will form the basis for the elicitation in sections 6.3 and 6.4 of this chapter. Stage one is the *setup* stage, which consists of identifying what aspects of the problem to elicit. This has been carried out to an extent in chapter 5.4.1 and 5.4.2 where the Bayesian Log-linear and Data Assessment Models, including priors, are specified. Consequently, the first step in elicitation is identifying exactly which element of data assessment is under consideration and which parameter in the model the prior relates to. Having identified, from chapter 5, the parameter for which the prior is specified it is also now clear what probability distribution the prior will take - a normal or gamma distribution, for example.

In stage two we *elicit* specific summaries of the subjective probability distributions for each parameter. In practical terms this means assigning a value for each hyper-parameter of each respective probability distribution. For example, for a normally distributed prior, this could be carried out by assessing the available evidence on the size of and uncertainty around the effect of the distortion of true flow for a specific parameter, for a given UK immgration data source. This would then provide a value for the mean and precision hyper-parameters, again specified in chapter 5.4.1 and 5.4.2.

Following this, stage three is to *fit* a probability distribution to these summaries. As Garthwaite et al (ibid) make clear, stages two and three are often carried out concurrently. This is the case in sections 6.3.2 and 6.4.2 of this chapter as the elicitation is designed with providing the relevant hyperparameters in mind.

Finally, stage 4, involves *assessing* the adequacy of the elicitation. When external expert judgement is being sought, there is the option of returning to the second stage –

elicitation – and obtaining more summaries from the expert. However, in this research, the expert judgement is being elicited from all appropriate available evidence, so this stage will be carried out in the sensitivity analysis.

At face value, the four stages seem relatively straight forward and simple. However, there are a range of considerations and assumptions upon which elicitation is based that need to be made clear. Firstly, it cannot be said that the expert judgements being elicited are the truth in an objective sense. Indeed, a subjectivist approach is being taken in this research and any claim that the models specified and priors elicited are some kind of 'objective truth' is misguided. Simply, the elicitation and the use of subjective priors should be viewed as part of the process of statistical modelling. It is ultimately part of developing our understanding of the uncertainty inherent in UK immigration statistics, based upon the available evidence. If new evidence came to light, the priors (and if appropriate, the model frameworks) could and would probably change.

In the same way that judgements and decisions are made in the specification of a statistical model, similar judgements need to be considered in the elicitation process. For example, taking care to avoid bias in the various judgements is a key consideration of any elicitation. One potential source of bias could arise from the desire to produce 'better estimates' or 'accurate estimates' of immigration. This could manifest itself in overstating the certainty of our prior judgements. To avoid this, priors, where possible, are elicited from available evidence that corroborates the general nature of the data assessments summarised in table 3.1. Furthermore, overstating the certainty of the prior judgements will simply prevent us addressing the second main aim of the research – to better understand the uncertainty inherent in the sources of data available to estimate UK immigration.

For the elicitation to be done well, it is important to distinguish between the quality of the expert knowledge and the accuracy with which that the knowledge is translated into probabilistic form in the mode of a prior probability distribution (Garthwaite et al 2005). A precise elicitation of an expert belief is extremely difficult, however, and might not be deemed worthwhile. A reasonable goal, though, of prior elicitation is to capture the main thrust of the expert opinion (ibid). With regard to the aims of this research this justification of prior elicitation rings true. The aim of the research is not to come up with a correct point estimate of true flow; as, in the absence of collecting far better data, this is unrealistic. Rather, it is use statistical models to both make better use of the evidence available and to better understand the amount of and specific sources of uncertainty inherent in the data

being used. Eliciting priors that clearly steer us to understanding better the distortion of true flow of immigration, in all the available data, is a contribution towards these two main aims of the research.

A further consideration is the difficulty of eliciting certain statistics. For example, in determining a prior distribution, in many applications, the variance of an unknown quantity, a parameter to be estimated, needs to be elicited. However, estimating the variance itself is problematic; as variance is a difficult concept to interpret and is often poorly estimated during elicitations (Beach and Scopp 1967 and Lathrop 1967, both cited in Garthwaite et al 2005). Garthwaite et al (ibid, page 16) suggest that one way of eliciting variances, without having to do it directly, is to elicit credible intervals, where the variance can be calculated based on the distributional assumptions made for that particular prior. Consequently, credible intervals, from which the variance can be calculated, are used in sections 6.3 and 6.4.

In terms of assessing the adequacy of elicited priors, as previously mentioned it is not helpful to assume that there is a true prior distribution that can be perfectly specified. The subjective assessments of the data available in this research are only based on available evidence, and could change dependent on additional evidence coming to light. Additionally, there are usually many distributions which fit the elicited summary values equally well. It could be the case that the posterior is insensitive to the choice of the fitted distribution and that other plausible distributions that fit the summaries would lead to essentially the same conclusions (O'Hagan 2005).

Furthermore, prior knowledge is inherently difficult to elicit and not easy to quantify without careful thought (Winkler 1967). If the elicitation was carried out in a different way – through the use of a panel of experts rather than from the available evidence, for example – then it is possible that the priors would be different for a given parameter and data source. Consequently, it is important to assess the impact of specifying different prior probability distributions for both the Bayesian Log-linear model and the Data Assessment Model. The impact of this becomes clear when the results of a sensitivity analysis, where the hyper-parameters of the prior probability distributions are allowed vary, are analysed in chapter 7.

One obvious source of uncertainty in elicitation is the fact that subjective judgements made by experts usually differ and will inevitably be rounded. This is on top of

the uncertainty already mentioned with regard to potential biases and the difficulty of eliciting precise summaries (O'Hagan 2005). Quantifying the effect of these uncertainties in elicitation and whether they matter can be determined in a sensitivity analysis. For example, if the posterior is sensitive to a certain prior then, with regard to the recommendations this research can make, in future work it will be important to either spend more time improving elicitation of this specific prior or to collect data where the specific distortion of true flow is less marked.

All of this section, so far, has referred to elicitation in univariate contexts, i.e. for single parameters. However, it is often the case that priors need to be elicited for a multivariate problem, such as this research. There is much less research on elicitations for applications of this kind than in a univariate context. If the variables are independent, though, this makes elicitation far more straight-forward as marginal probabilities for each variable can be elicited (O'Hagan 2005). Independence in the context of prior elicitation effectively means that the knowledge gained for one parameter does not affect the knowledge held for the other parameters in the model. This is an assumption that is made for both models, specified in chapter 5.

6.2.2 Applications in Demographic Models

The application of Bayesian methods to address some of the known problems in data availability and quality and definitional differences was suggested and considered in Willekens' (1994) paper. Prior to this, Cooke (1991) outlines how one can treat assessments of data sets - from what is known about data collection, for example - as data themselves. Willekens (1994) outlines some of the key considerations one needs to make when incorporating expert knowledge about data, into a model, as data itself. This discussion moves beyond the combining of data, which is addressed in Chapter 4, and on to the addition of judgemental data into the estimation process. Often, among experts and users of population statistics, there is useful and potentially useable knowledge, which is under-utilised. This knowledge is often too imprecise or fragmentary for it to be included in a formal data base or model (ibid: page 18). This is problematic and far from perfect; but, the key consideration that the researcher needs to make here is whether the somewhat imprecise and uncertain judgemental data could potentially improve the predictive value of the data already being used in the formal model (McNees 1990 cited in Willekens 1994).

There is a limited literature on the use of expert judgement in demographic statistical models to estimate migration. Bijak and Wiśniowski (2010) carry out an elicitation of subjective judgement from a panel of experts, using a Delphi survey approach, for migration forecasts. This is extended by Wiśniowski, Bijak and Shang (2014), who make use of a 'roulette betting approach' in their paper on forecasting Scottish migration following the independence referendum. Experts are asked to make a judgement about whether selected components of migration were to increase, decrease or stay the same. Following this, to elicit a level of uncertainty, they were asked to state how much they would be willing to 'bet' on the level of migration in 2021 (ibid, page 458). The betting took the form of placing a number of 'chips', in a roulette style, on given range of the future level of migration.

Wheldon et al (2013) use a Bayesian approach to reconstruct past populations where there are fragmentary data and taking into account the measurement error of each of the sources. Informative priors are elicited for the measurement errors for vital rates, migration rates and population counts at baseline. For each country, expert opinions were sought and the priors were set in relation to each of the aforementioned demographic parameters. As touched upon is section 6.2.1, Wheldon et al (2013) make clear that even though exact prior knowledge amongst the experts of the respective hyperparameters of the priors is unlikely, one can still reasonably derive some useful expert knowledge of the data sources (Wheldon et al 2013, page 99). A straightforward method to elicit expert knowledge is used, where plausible ranges for the mean absolute error are specified by the experts for each given demographic parameter.

Finally work carried out by Raymer et al (2013) on The Integrated Model of European Migration (IMEM) used expert opinion to harmonise and take into account inconsistencies and differing levels of accuracy in multiple sources of international migration data. In this section, the main and relevant contributions from Wiśniowski et al's (2013) paper on prior elicitation for the IMEM model are outlined and used to help inform the prior elicitation in sections 6.3 and 6.4. Similar to the UK Data Assessment Model specified in section 5.4.2 Wiśniowski et al require prior information on the quality of the data sources and differences in the various aspects of data measurement characteristics. To achieve this external expert judgement was sought. Being a multi-source problem, experts were asked to rate the credibility they gave to different types of collections mechanisms, such as population registers and surveys, and to compare sending and receiving countries'

data. They were also asked about any biases in the reported migration flows (Wiśniowski et al 2013, page 587). These opinions from the panel of experts were converted into probability distributions which were then subsequently combined into a single set of distributions.

To facilitate these expert judgements a two-stage Delphi framework was employed. Experts opinions, in this process, were allowed to be informed and influenced by other views of experts from the panel, thus, facilitating an exchange of opinion and views and clarification of any underlying concepts and ideas. In the design of the Delphi survey, Wiśniowski et al (2013) carefully try to avoid respondents from being overconfident in the expression of their opinions.

One of the key contributions of the survey approach used by Wiśniowski et al (2013) was getting the experts to think of migration collection systems in general, rather than specific country experiences. Eliciting subjective judgements on parameters for many individual countries is problematic and is expanded upon in section 6.3, where priors are elicited for the Bayesian Log-linear model of UK immigration.

One section of the questionnaire concerned the duration of stay criteria included in the UN definition of migration. The experts were asked to consider how different timing criteria used by different countries might affect the relative levels of reported migration (Wiśniowski et al 2013, page 590). In the Data Assessment Model of UK immigration, specified in section 5.4.2, the 'definition' parameter takes into account how each source of data distorts true flow with regard to duration of stay. As such, the expert opinion elicited from the Delphi survey, on duration of stay, is one source of evidence that could be used in elicitation for this research (see section 6.4.3).

The most challenging parameter to assess was the overall accuracy of a given data collection system (Wiśniowski et al 2013). This is certainly the case with the UK immigration data. A further contribution of prior elicitation by Wiśniowski et al that is applied in this research is that they only sought judgements on the data and measurement aspects of the underlying migration flows (ibid).

Many problems need to be addressed and taken into account before the judgemental data from chapter 3 can be used in a model for prediction. Separating useable expert knowledge from information that is contained in sources of data is a key consideration made by Willekens (1994). With regard to this research, this is problematic;

keeping 'expert information' and information on data collection separate from the data available for UK immigration estimation, having already undertaken an assessment of the data in chapter 3 and a log-linear analysis in chapter 4, is difficult. The second problem that needs to be addressed, according to Willekens (ibid), are the issues of bias and uncertainty in the expert knowledge being used.

Keeping expert information independent of the data and the formalisation of this information is also considered and addressed in a paper on forecasting international migration by Bijak and Wiśniowski (2010) through the use of a multi-stage survey of international migration experts. The survey-design means that generalised judgements of international migration data are obtained iteratively, with the respondents receiving aggregated feedback of the responses at each stage.

Wiśniowski et al (2013) ensure the data elicited from expert opinion is independent from the data in the model through addressing specific characteristics of data collection and definitions. For example, they make clear in their surveys that they want an assessment of the extent to which specific measurements of international migration deviate from a given benchmark; the UN definition of long term migration and how certain each of the experts are about this. This approach – consideration of to what extent true flow is distorted and the certainty with which this prior judgment can be expressed – is used in the prior elicitation in this chapter.

From this review it is clear that in the remaining sections of this chapter that the following needs to be considered. First, prior information needs to be separate from the data used in the model of migration; reusing data in both the prior and likelihood is to be avoided. Second, expert judgement needs to be specified in a clear and coherent way, which lends itself to being specified as a probability distribution. As such, some kind of consideration of uncertainty needs to be explicitly included in the judgement. Third, the specification of uncertainty in the priors needs to explicitly consider the distribution being used.

6.3 Prior Elicitation for Bayesian Log-linear Model of UK Immigration

The purpose of this section is to elicit prior probability distributions for the main effects parameters of the Bayesian Log-linear model of UK immigration. In a broad sense

the motivation for the prior elicitation in this chapter is the inclusion of further sources of data which are yet to have been used in the estimation of true flow. The two sources of auxiliary data that are utilised through the priors for the Bayesian Log-linear model are the 2001 Census estimate of immigration to the UK, and the Flag 4 GP Registration data of foreign nationals. The full model is specified in chapter 5.4.1. The priors placed on the main effects terms in the model act as quasi-alternative marginal effects. The uncertainty elicited will be a reflection of our confidence in the Census and Flag 4 data as alternative marginal effects in the estimation of UK immigration via a log-linear framework.

Up to this point the log-linear model has constrained the estimates to the more reliable margins of the IPS and then used the HESA and DWP data to provide the detailed pattern of immigration over time for individual citizenships. Through adopting a Bayesian approach we can temper the above with the addition of priors, which are centred on further auxiliary data identified, and are assessed in relation to true flow.

As outlined in chapter 5.4.1, the 2001 Census data and Flag 4 GP registration data are used to set the prior means for the main effects parameters. This is also referred to as centring the priors on the auxiliary data sources – census and Flag 4 respectively. However, with the priors being specified in relation to the overall and main effects parameters, simply centring the priors on the particular value of the auxiliary data in question would lead to a mis-specification of the priors. As such, it is necessary to outline in detail the parameters of interest for which priors are elicited. A general approach to eliciting the mean values for informative priors for the main effects of a Bayesian log-linear model of a contingency table is proposed in section 6.3.1. This section also provides a thorough relevant explanation and interpretation of the parameters for which priors are elicited.

In section 6.3.2 there is an account of the elicitation of the prior hyperparameters for each of the parameters in the model. At the beginning of this section, however, a method is proposed for the elicitation of the uncertainty of a prior, which uses the quantiles in the tail of each distribution. Effectively a wide credible interval for the parameter of interest is elicited and then used to calculate the uncertainty. This method is applied in both the prior elicitation for the Bayesian Log-linear model in section 6.3.2 and for the elicitation of the priors for the Data Assessment Model in section 6.4.2.

Finally, in section 6.3.3 there is a discussion of the prior elicitation of the main effects. The main contributions and the limitations of the elicitation process are outlined here.

6.3.1 Parameters of Interest – Overall and Main Effects

As outlined in section 6.2.1, the first step in prior elicitation is identifying exactly what is being elicited. This is referred to as the *setup* stage by Garthwaite et al (2005). There are two elements to this for the Bayesian Log-linear model. The first is identifying the parameters and zeroing in on their interpretation so that the prior elicitation directly relates to the overall effects and main effects of the likelihood parts of the student and non-student models. The second is linking the identification of the parameter and subsequent prior elicitation to the appropriate qualitative assessment of the data in relation to true flow. This is especially important in the elicitation of priors for the Bayesian Log-linear model, as the assessment of the distortion of true flow takes place in the context of a log-linear framework.

The student and non-student models are specified in section 5.4.1 by equations 5.9 and 5.10. Priors for the overall and main effects in the log-linear likelihood are specified below. Each of the priors are assumed to take a normal distribution, $N(\theta, \tau)$ where θ is the prior mean and τ is the precision. As such, mean and precision hyperparameters need to be elicited for each parameter in the student and non-student models respectively.

The priors for the overall effects are denoted by $\lambda^S \sim N\left(\theta_{\lambda^S}, \tau_{\lambda^S}\right)$ and $\lambda^W \sim N(\theta_{\lambda^W}, \tau_{\lambda^W})$ for students and non-students respectively, where θ_{λ^S} and θ_{λ^W} are the prior means for the student and non-student overall effects and τ_{λ^S} and τ_{λ^W} are the precision hyperparameters for students and non-students. The subscripts used for the prior hyperparameters denote the parameter for which the prior is being set.

Precision is the inverse of variance and, as such, the higher the level of precision elicited the lower the variance, which leads to a tighter probability distribution around the mean. Consequently if we have a large amount of confidence in the prior a high precision will be elicited and vice versa for a prior where there are low levels of confidence.

The main effects priors for both the student and non-student model are also assumed to take independent normal distributions, they are denoted below:

$$\lambda_i^S \sim N(\theta_{\lambda_i^S}, \tau_{\lambda_i^S}) \quad \lambda_t^S \sim N(\theta_{\lambda_t^S}, \tau_{\lambda_t^S}) \qquad \lambda_i^W \sim N(\theta_{\lambda_i^W}, \tau_{\lambda_i^W}) \quad \lambda_t^W \sim N(\theta_{\lambda_t^W}, \tau_{\lambda_t^W}), \quad (6.1)$$

where each of the subscripts for the mean and precision hyperparameters denote the parameter of the likelihood, that the prior is specified for. Simply centring the priors on

just the value taken from the auxiliary data that is being used is not adequate; we need to refer back to the likelihood model to understand how the gap is bridged between the data being used as a prior and the prior's role in the full probability model.

The Bayesian Log-linear model is specified using the corner cell constraint where, in this case, $\mu_{11,9}$ is estimated from the overall effect λ. Each of the main effects parameters in the independence model are, therefore, interpreted in relation to a reference category. Consequently, with the model constrained to the final corner cell, for λ_i the reference category is 'all-other' and for λ_t the reference category is 2010, for both the student and the non-student models. Take note, when there is no superscript indicating whether discussion relates to the student or the non-student model, reference is being made to both models to save on repetition of notation.

The constraint of the model outlined above means that the differences between two parameters for a given variable relate to the log odds of making one response, relative to the other, on that variable (Agresti 2013). Strictly speaking, this means that auxiliary information about these two references cannot be included neatly in the form of two prior distributions for 'all-other' and '2010' main effects. Any consideration of these two categories, in a prior judgement, has to be elicited for the main effects $\lambda^S$and λ^W.

The inclusion of an offset term in the likelihood, however, complicates interpretation of the parameters slightly. To aid interpretation of the main effects parameters, and thus guide prior elicitation, the log-linear model with offset term, can be simply rearranged as follows:

$$\log \mu_{it} = \lambda + \lambda_i + \lambda_t + \log(n_{it})$$

$$\log \frac{\mu_{it}}{n_{it}} = \lambda + \lambda_i + \lambda_t \qquad (6.2)$$

where μ_{it} denotes the fitted values of the independence model and n_{it} is the offset term. The parameters can now be treated as linear predictors of a ratio of fitted values to the offset term. They are also, conveniently with regard to understanding how the priors need to be set, the only terms on the right-hand side of the equation.

With regard to the prior elicitation for the citizenship-specific parameter, λ_i , auxiliary information from the 2001 census is used in the centring of the distribution. The

2001 census estimate of immigration is the only remaining source of data, that has not been used in the specification of the likelihood, that can used as a prior for this particular marginal effect. So, firstly the estimation of the rows of the contingency table, denoted by subscript *i,* is considered.

Conceptually, the census data itself, is equivalent to μ_i, as it is an auxiliary source of data which is being used as a prior judgement on the predictor of the total immigration of a given citizenship, over the period 2002-2010. Having rearranged the likelihood model (equation 6.2) and for the purposes of elicitation, if the census data denoted as c_i is, as already mentioned, equivalent to μ_i then

$$\log\frac{c_i}{n_i} \approx \log\frac{\mu_i}{n_i} \ , \tag{6.3}$$

where, and throughout this section, the notation ≈denotes 'conceptually equivalent'. So, considering just the parameters specified for the marginal probabilities of *i,* disregarding the time-specific effects and substituting $\log\frac{c_i}{n_i}$ for $\log\frac{\mu_i}{n_i}$, the likelihood model can be further rearranged to give

$$\log\left(\frac{c_i}{n_i}\right) - \lambda \approx \lambda_i \tag{6.4}$$

where the ratio of the census to the offset measurements for the reference year, for a given citizenship, minus the overall effect, is equivalent to the citizenship-specific main effect. As discussed in chapter 4, and in more detail above, one of the main-effect parameters is redundant, as each parameter is set in relation to a reference category. In this case for λ_i the reference category is the 'all-other' citizenship category. As previously mentioned the reference category is estimated from the overall effect, λ. So, with this in mind 6.4 can be re-written as follows:

$$\log\left(\frac{c_i}{n_i}\right) - \log\left(\frac{c_{ref}}{n_{ref}}\right) \approx \lambda_i \ , \tag{6.5}$$

which is equivalent to

$$\log\left(\frac{c_i/c_{ref}}{n_i/n_{ref}}\right) \approx \lambda_i \ , \tag{6.6}$$

where subscript ref denotes the citizenship reference category, 'all-other'. Performing the calculation, set out by equation 6.6, for each citizenship, gives the mean value for the citizenship-effect priors specified for the λ_i main effect.

To set the prior means for the column main effect λ_t , where subscript *t* denotes year, a similar calculation to that specified by equation 6.6 is made for each year of interest. The auxiliary data used in the calculation to centre the priors for λ_t is the Flag 4 GP registration data. A full description and assessment of this source of data is outlined in chapter 3. The Flag 4 data is denoted as f_t and the calculation for the prior means ultimately takes the form

$$\log\left(\frac{f_t/f_{ref}}{n_t/n_{*ref}}\right) \approx \lambda_t \ . \tag{6.7}$$

where subscript $*ref$ denotes the reference category for the offset term. The calculations, specified by equations 6.6 and 6.7 are used to set the prior means. In terms of model specification, the prior means are now equivalent to the mean of main effects parameters of the likelihood. Essentially they can be interpreted in a similar way. The calculations to centre the priors for λ_i and λ_t is carried out for both the student and non-student models, as priors are specified for the student and non-student models.

The same Census and Flag 4 auxiliary data is used for both the student and non-student parts of the Bayesian Log-linear model, which is far from perfect, as ideally a different source of auxiliary data would be available for each. However, this limitation is addressed to a certain extent later in the prior elicitation, as different levels of certainty are elicited for students and non-students for each of the auxiliary sources of data.

For the likelihood part of the student probability model (see chapter 5.4.1), HESA data is used as an offset term and in the non-student model the DWP data is used as an offset term. As such, even though the prior means for the student and non-student model are elicited using the same auxiliary data, their values are different for the student and non-student models. This is detailed, in full, in the following section, 6.3.2.

Interpretation of the prior means calculated by equations 6.6 and 6.7 needs to be considered. If the prior mean is greater than one for a given citizenship or year, it means that compared to the reference category there is an overcount in auxiliary data in relation to

the offset. The reverse is true if the prior mean is negative. These values for the prior means can then be compared to the parameter estimates of the frequentist log-linear model. The difference between these two values can be attributed to either the auxiliary data prior means being a better or worse reflection of true flow, for the main effects of a log-linear model, than the IPS parameters.

Consequently, the prior means are calculated in the following section, where they are compared to the values for the log-linear parameters. For each category of the main effects parameters for the student and non-student models, an uncertainty is elicited which captures the confidence we have in the auxiliary data sources as alternative marginal constraints.

6.3.2 Elicitation of Hyperparameters

Up until this point in the study, interpreting the parameters of the models specified within a log-linear framework has not been of concern. The focus has been on the fitted values of the models; and how the use of an offset within a log-linear framework has utilised the strength of IPS and HESA and DWP data respectively to estimate immigration. However, with the requirement that the prior means are set to correspond with the main effects, interpretation of the parameters has become necessary – if one is to elicit a level of uncertainty for a prior then one needs to know exactly what this uncertainty is being ascribed to. As such, the level of uncertainty for each prior is directly linked to the interpretation of the parameters in the model.

Interpreting what the main effects priors mean in terms of estimating true flow, within a log-linear framework, is not straightforward, however. As outlined by equations 6.6 and 6.7 the priors are centred through the use of auxiliary data and as such they are not a simple assessment of how well the data used in the model estimates true flow. Rather, their uncertainty is an assessment of the confidence we have in the auxiliary data sources as alternative marginal effects within a log-linear framework. To elicit a level of uncertainty precisely, that directly relates to this is not realistic. It is too far removed from an everyday understanding of how the characteristics of data collection, for the auxiliary sources used to centre the priors, has distorted true flow.

In this section the second and third stages of prior elicitation, outlined in section 6.2.1, is carried out. These stages, as proposed by Garthwaite et al (2005) are the *elicitation* of the quantities of interest and the *fitting* of the probability distribution that is used as prior in the model.

The general approach to prior elicitation is to specify indicative levels of high, low and medium levels of certainty for the hyperparameters. These can then be varied in a sensitivity analysis to determine the level of certainty that is required for the posterior to be influenced by the auxiliary sources of data. Effectively, the use of the census and Flag 4 GP registration data to set the prior means is specifying a prior belief about the efficacy and suitability of these two sources of auxiliary data in estimating the marginal probabilities of true flow.

As a result, the uncertainty for each of the priors is elicited for each value of *i* and *t* for each of the τ hyperparameters in a variety of ways. Four different full probability models are estimated and a summary of the full elicitation can be found in table 6.1, for students and non-students, later in this section. Superscripts [1], [2], [3] and [4] denote which of the four precision levels are being discussed. The prior means remain the same for all four estimates of the model, as they are calculated using the auxiliary census and flag 4 data. The results of the calculations using equations 6.6 and 6.7 to centre the priors, are detailed in table 6.1. In general, the level of certainty expressed for the level of τ is increased for each run of the model, with the precision level [3] and precision [4] set of priors including country specific levels for $\tau_{\lambda_i^W}$.

Firstly, however, the priors for the overall effects are elicited. The overall effects in the Bayesian likelihood provide the estimates for the reference categories, all-other and 2010 respectively as the model is constrained to the corner cell of the contingency table. Providing a prior median and precision for the parameter, which predicts the fitted values for $\hat{y}^S_{11,9}$ and $\hat{y}^W_{11,9}$, is problematic, as there is no auxiliary data that can be applied. Consequently, the priors for the overall effects are elicited as being vague, $\lambda^S \sim N(0, 0.0001)$ and $\lambda^W \sim N(0, 0.0001)$. As result the posterior of the main effect closely reflects the IPS-based likelihood. The impact of this on all the estimates of true flow is discussed further in chapter 7.

Parameter	Student Priors					Non Student Priors				
	Mean	Prec [1]	Prec [2]	Prec [3]	Prec [4]	Mean	Prec [1]	Prec [2]	Prec [3]	Prec [4]
Overall Effect	0	0.0001	0.0001	0.0001	0.0001	0	0.0001	0.0001	0.0001	0.0001
Australia	2.83	1	1	1	10	0.73	10	10	100	1000
China	-2.25	1	1	1	10	-0.46	10	10	1	10
France	0.31	1	1	1	10	0.54	10	10	10	100
Germany	0.65	1	1	1	10	1.15	10	10	10	100
India	-0.63	1	1	1	10	-0.82	10	10	10	100
Pakistan	-0.06	1	1	1	10	-0.71	10	10	100	1000
Philippines	1.45	1	1	1	10	-0.17	10	10	10	100
Poland	-0.86	1	1	1	10	-3.52	10	10	1	10
S. Africa	2.66	1	1	1	10	0.69	10	10	100	1000
USA	0.80	1	1	1	10	1.78	10	10	100	1000
2002	0.27	10	100	10	1000	0.39	100	1000	1000	10000
2003	0.15	10	100	10	1000	0.30	100	1000	1000	10000
2004	0.11	10	100	10	1000	0.21	100	1000	1000	10000
2005	0.18	10	100	10	1000	-0.07	100	1000	1000	10000
2006	0.24	10	100	10	1000	-0.04	100	1000	1000	10000
2007	0.20	10	100	10	1000	-0.22	100	1000	1000	10000
2008	0.18	10	100	10	1000	-0.03	100	1000	1000	10000
2009	0.05	10	100	10	1000	0.04	100	1000	1000	10000

Table 6.1: Summary of Prior Elicitation for the Bayesian Log-linear Model

Below, there is an assessment of how good an alternative set of true flow marginal constraints each source of auxiliary data is. This is carried out for both the student and non-student models respectively. It is important to note, however, how complete each of the sources of auxiliary data are as quasi marginal effects. The 2001 Census only provides an estimate of immigration for one year. Additionally this year, 2001, is outside of the study period of 2002-2010. An assumption that has to be made, therefore, is that the relative average effect across all years for each of the citizenships is the similar to prior medians elicited from just the 2001 census data. This is not a problem with the Flag 4 data, which is being used an estimate for the total of all citizenships for each individual year.

Consequently, as a rule of thumb, the Flag 4 priors are elicited with a higher level of certainty than the Census priors.

For the first two sets of model estimates, the precision of each of the τ hyperparameters is assumed to be the same for each level of *i* of *t* resulting in four different levels of certainty being elicited (see table 6.1). To elicit the precision hyperparameters, we need to refer back to the data assessment of Census and Flag 4 data in chapter 3.

To begin with, the Census data for students is considered for the $\tau^{[1]}_{\lambda_i^S}$ precision hyperparameter. The relevant data assessment criteria which are relevant are coverage and bias. The detail in table 3.1, from the data assessment of chapter 3, indicates that there is a lack of coverage for students from Scotland and Northern Ireland. If the lack of coverage affects each of the top 10 citizenships in the same way, and the level of non-UK domiciled students as a proportion of all HE students is the same in England and Wales as it is in Scotland and Northern Ireland, then the level of uncertainty should not be affected. However, there is no concrete evidence that this is, or is not the case. So, the judgement for coverage is that we should treat the census data for students with caution.

With regard to bias, students could be less likely to fill out the census form and are therefore seen as hard to count. In fact, in a case study report on the collection of data for the 2001 census in Westminster constituency, produced by the Statistics Commission (which was replaced by the UK Statistics Authority), it is outlined that students were particularly hard to count. Enumerators failed to gain access to halls of residence, for instance, to capture people who did not fill out their Census form voluntarily (Statistics Commission 2004). As a result, the distortion of true flow for the Census data used for the priors in the student model is deemed to be large. Consequently for all values of *i* (the top 10 citizenships, as the 'all-other' citizenship is set as the reference category) the precision for $\tau^{[1]}_{\lambda_i^S}$ is set to be low, to take an assumed value of 1.

As previously mentioned, the Census data is being used to help centre the priors for both students and non-students. For non-students the Census data is a relatively better reflection of true flow for the marginal constraints of the model. There is still the same problem with coverage; however, even though non-student migrants are relatively hard to count, the judgement is that they are more likely to respond to Census questionnaire or to be living in accommodation that is more accessible to enumeration if there is non-

response. So for all levels of *i* the precision for $\tau^{[1]}_{\lambda_i^W}$ is still set as being low, but there is more certainty here than in the student model. As such it takes a value of 10.

For the values of $\tau^{[1]}_{\lambda_t^S}$ and $\tau^{[1]}_{\lambda_t^W}$ the Flag 4 GP registration data needs to be assessed in relation to true flow. Two precision levels are elicited, one for students and one for non-students, which take the same values across all levels of *t* for the years 2002-2009 (note that 2010 is the reference category for *t* and is estimated from the overall effect).

With regard to coverage, Flag 4 data only covers England and Wales. This should not be a problem if the proportion of immigrants registering at the GP is the same in England and Wales as it is in Scotland and Northern Ireland. For definition, there is no duration of stay criteria for GP registrations, which means that short term and circular migrants could be included in this data set. This is not necessarily a distortion of true flow for the priors, as, if the proportion of short term migrants registering at their GPs is the same across each of the values of *t,* then the precision can be set the same for each year.

Considering the bias criteria, it is documented, that young males are less likely to register with a doctor (Raymer et al 2012, Stagg et al 2012) and that some migrants are only likely to register with a doctor if they require medical assistance (Greater London Authority 2010). Seeing as though one would expect most students to be aged between 18 and 24, it is probable that the flag 4 data undercounts the number of students who migrate to the UK each year. Thinking about the interaction of the bias and definition criteria, lots short term student migrants are unlikely to register with a GP. There is no strong evidence of this either way, though.

The overall Flag 4 GP registration figure increases over time from 429,752 in 2002 to 595,341 in 2009. For the precision to change, over time, there would have to be a changing number of students migrating to the UK who were either more or less likely to register with a doctor. With the expansion of freedom of movement in 2004, however, it is likely that there was a large short-term in-flow including seasonal migration. It could be argued that as result of this, in the years following 2004 the level of non-student precision needs to be set lower than for the preceding years. However, from the assessment of data in chapter 3, we know that shorter-term migrants, who are slightly more likely to be male especially if they arrived after 2004, are less likely to register at their GP. Again, though, there is no strong evidence of this that can be used directly in the elicitation of the precision term.

With regard to the bias criterion, the rate of GP registration for all immigrants is probably lower for students than it is for non-students as it could be argued that non-students are more likely to register with a doctor. Additionally, non-students are more likely to migrate to the UK as part of a family group and with females more likely to register with a GP (Stagg et al 2012), it is probable that a greater number of non-student males are registered to their GPs through registering as a family. Even taking this into account though, there will almost certainly be an under-registration of long term migrants with GPs (see chapter 3 for a full discussion). As a result, taking the above into account we have more confidence in the Flag 4 data for non-student priors. Consequently, for the first set of model estimates the following values of the precision parameter are assumed $\tau^{[1]}_{\lambda_t^S} = 10$ and $\tau^{[1]}_{\lambda_t^W} = 100$.

With the Flag 4 data being deemed a better reflection of the true flow marginal effects for non-students, for the second set of model estimates, the precision level for census priors is kept at $\tau^{[2]}_{\lambda_i^S} = 1$ and $\tau^{[2]}_{\lambda_i^W} = 10$, whereas the precision for Flag 4 priors is increased by one order of magnitude to $\tau^{[2]}_{\lambda_t^S} = 100$ and $\tau^{[2]}_{\lambda_t^W} = 1000$.

For the third and fourth set of precision levels, country-specific priors are elicited for the precision hyperparamaters for the non-student main effect, λ_i^W. These are detailed in table 6.1. The precision level [3] values are detailed below.

For the priors specified for λ_i^S the precision is kept at $\tau^{[3]}_{\lambda_i^S} = 1$, as our confidence in the Census for the student marginal true-flow-effect, remains low. For the Flag 4 data, the precision values for students revert back to $\tau^{[3]}_{\lambda_t^S} = 10$.

The values of $\tau^{[3]}_{\lambda_i^W}$ are elicited as follows. The top 10 flows for the λ_i^W main effect are split into three relative groups – good, medium and poor precision. The citizenships that are in the good category are deemed to have a relatively close match to marginal true flow and the precision term is set at 100. These citizenships are Australia, Pakistan, South Africa and USA. From chapter 2 we know that each of these citizenships have well-established flows of immigration to the UK. This mitigates the limitation that the 2001 census data is just a snapshot of immigration being used to a prior for a main effect across multiple years. More speculatively, and with regard to data collection, there are well-

established migrant communities for each of these citizenships, as a result there might be a better census completion rate.

France, Germany, India and the Philippines are placed in the medium precision group and have a precision value set at 10. For India, there is a sizeable flow of international students included in census data, otherwise as we know from chapter 2 this flow is relatively stable over time. France and Germany's flows are also made up of relatively significant student inflow, however, these inflows are also relatively well-established with a large French population in London, for example (ONS 2012 c).

Finally, the poor precision group includes China and Poland. For China, there has been an increasing number of students migrating to the UK. So for Chinese non-students it could be argued that there is overcount in the Census data. The Census data is taken from 2001, though, and it could be the case that there was only low Chinese student immigration back then. Up until this point, direct assessment as the quantity of the estimate has been avoided, to ensure that the information elicited for the priors is kept separate from the likelihood. However, the 2001 Census estimate of annual immigration from China is very low – 5,675, which is around seven times lower than for the USA - and cannot be ignored in the prior elicitation. As such, and taking into the uncertainty around the inclusion of students in the Census estimate, a judgment is made to set the precision to 1. For Poland the stand-out reason to elicit a high level of uncertainty, is the fact that the data is from 2001 and we know that from 2004 onwards there has relatively been much higher levels of Polish immigration to the UK. For the citizenship main effect to be a good reflection of true flow it needs to take into account the average level of immigration for given level of i for all values for t.

For the fourth and final set of model estimates, all levels of precision are increased to test the sensitivity of the posterior to priors with relatively high levels of certainty. The values for the precision hyperparameters, in relation to each other, are still a reflection of the overall judgements outlined throughout this section, it just that they have all been elicited with a high-level of precision. All values of the elicited hyperparameters, for the four models, are summarised in table 6.1.

6.3.3 Discussion and Limitations

In this section a method has been proposed to elicit informative prior medians, of normally distributed priors, for a Bayesian log-linear model of contingency tables, when one has an alternative source of marginal constraints. Examples of applications of Bayesian log-linear models within the literature, which make use of informative priors, are rather limited. As such a contribution of this part of the research is to propose an approach that can be used to bridge the gap between the subjective information that their prior judgment is based on and how to elicit a prior median which corresponds to the parameter of interest.

Prior elicitation for the overall effects in the model remains problematic. One of the reasons for this is the corner cell constraint that is used in the model specification in the previous chapter. Elicitation for the corner cell was not possible, so a diffuse prior was specified. This means that for the posterior estimates of the reference categories – all-other and 2010 – the likelihood will be dominant. Future research could look at using different model constraints and readjusting the prior elicitation calculations for the prior medians and the interpretation of the uncertainty accordingly.

Furthermore, eliciting the uncertainty for each prior distribution has been problematic too. The level of the precision parameters τ is effectively a broad judgement of the quality of the auxiliary data as a marginal representation of true flow. The sensitivity analysis proposed, through four sets of model estimation, will give an indication of how sensitive the likelihood is to different elicitations of prior uncertainty.

6.4 Prior Elicitation for Data Assessment Model of UK Immigration

In this section the priors for the Data Assessment Model, specified in chapter 5.4.2, are elicited. The full probability model is specified in order to capture the distortions of true flow, through the framework of the data assessment criteria outlined in chapter 3. The priors, and the general probability distribution each assume, have already been specified in 5.4.2. As such, the main motivation for this section is the need to elicit values for each of the hyperparameters of the data assessment priors.

Each of the parameters of interest in the Data Assessment Model are specified specifically to capture both the nature and uncertainty of each of the data assessment

criteria outlined in chapter 3. As a result the priors can be closely based on our *a priori* knowledge of data collection, the concepts and definitions each of the sources are based on. One of the main challenges in the prior elicitation of the Bayesian Log-linear model in section 6.3 was bridging the gap between the main effects parameters and the subjective judgement of the auxiliary data set used to centre the priors. This is not a problem for the priors specified for the Data Assessment Model. However, because the model is not constrained to a log-linear framework the role of adequate priors is far more consequential for the Data Assessment Model. As such, a clear and considered process needs to be followed in the elicitation of the distortion of true flow and the judgements of the associated uncertainty.

In section 6.4.1 the parameters of interest and the hyperparameters that need to be elicited are identified. How each of these hyperparameters is set, including a consideration of dealing with parameters which are specified on the log-scale and different types of probability distributions, are outlined. Following this, in section 6.4.2 the hyperparameters for each of the priors are elicited. The evidence that each elicitation is based on is outlined in detail and the priors are detailed in full in table 6.2 at the end of the section.

Finally in section 6.4.3 there is a discussion of the elicitation for the Data Assessment model, including a reflection on the limitations of the process. Strategies for the sensitivity analysis, which is carried out in chapter 7, are briefly proposed.

However, firstly a general approach to prior elicitation, based on eliciting a credible interval using quantiles, is outlined. This method is then applied in the prior elicitation for the Data Assessment Model. As mentioned earlier in the chapter, for prior elicitation to be effective it is important it is carried out in an intuitive and understandable way. As specified in chapter 5.4.1, however, the IPS data for both students and non-students is assumed to be generated by a Poission process. Consequently, the priors need to be specified on the log-scale.

This, though, is not straight forward, as it is not intuitive to think of the uncertainty of immigration estimates on the log scale. Furthermore, to obtain the precision hyperparameter, one needs an estimate of the variance, as $\tau = \frac{1}{\sigma^2}$. This too, as outlined in section 6.2.1 is not straight forward, as the variance statistic is not intuitively relatable to substantive subject matter. Garthwaite et al (2005) propose considering an elicitation of

credible intervals, where the variance (and in turn precision) can be calculated based on the distributional assumptions made for that particular prior.

The quantiles of a given distribution are invariant to monotonous transformations, such as the logarithmic transformation required to set the priors for the log-linear likelihood parameters. A log transformation of the data preserves the order of the quantiles. So, one can elicit a quantile at either tale of the prior distribution of interest on the linear scale before applying a log transformation. This allows an elicitation of a credible interval, which is more readily interpretable than a variance statistic

Specifying summaries, based on quantiles, of the subjective assessment of the data and then fitting a particular distribution to the elicited summary statistics may seem crude, but there is basically no alternative (cf. O'Hagan 2005). Even though the assessments of the data sources, in relation to true flow, in this research are based on evidence where possible, quantifying our certainty of the distortion is subjective. Translating this subjective judgement into precise prior distributions is not as straight forward as measuring something like the length of a piece of wood (O'Hagan 2005). As touched upon in section 6.2.1, it is more important that priors capture the main thrust of the subjective judgement and additional evidence they contribute to the full probability model.

For a normally distributed prior, one can use an elicitation of a quantile in either tale of a distribution to calculate the precision hyperparameter. We know that 95% of the probability mass under the normal distribution curve lies within 1.96 standard deviations of the mean, which is equal to the median. Consequently to obtain the standard deviation, for the prior, one can use the one of the following equations:

$$\sigma = (\log \mu - \log q_l)/1.96 \tag{6.8}$$

$$\sigma = (\log q_h - \log \mu)/1.96 \tag{6.9}$$

where σ is the standard deviation, μ is the mean (and median), q_l is the value of the 2.5 percentile and q_h is the value of the 97.5 percentile. Both the median and either q_l or q_h are transformed to the log-scale after they have been elicited, to ensure an appropriate standard deviation is calculated in either equations 6.8 or 6.9. With the elicited credible intervals coming from either the 2.5 or the 97.5 percentile, one needs to state a value that has a 1/40 probability, for that parameter, to be a good reflection of true flow. Having calculated the deviation one can then obtain the precision as follows: $\tau = \frac{1}{\theta^2}$.

6.4.1 Parameters of Interest – Data Assessment Parameters

As outlined in section 6.4.2 the first step in prior elicitation is the *setup* stage (Garthwaite et al 2005). In this section the parameters of interest, for which elicitation needs to be carried out, are identified and discussed. Furthermore, given the probabilities that have been specified in chapter 5.4.2 there is also a discussion and explanation about how each of the hyperparameters is calculated.

Each of the priors are specified in chapter 5.4.2 and are specified to explicitly capture the nature and uncertainty for the judgements of true flow distortion, detailed in chapter 3. There are three different types of prior distributions specified here – normal, gamma and log-normal distributions. Table 6.2 in the next section has details which parameters are specified with which probability distribution. The parameterisation of the normal distribution is the same as in section 6.3, $N(\theta, \tau)$ with hyperparameters θ and τ denoting the median (which is equivalent to the mean in a normal distribution) and the precision respectively. For the data assessment model the hyperparameter θ is referred to as the median, as the elicitation is based on the quantile method outlined previously.

As detailed in chapter 5.4.2, the overall distributions of the data assessment equations is log-normal, which means that each of the data assessment priors need to be elicited on the log-scale. The median and precision hyperparameters for each of the normally distributed priors are elicited by the quantile method, using equations 6.8 and 6.9.

Consequently, for normally-distributed priors, θ can be elicited directly from the median value of the prior, which is set according to available evidence or from subjective judgment of the distortion of true flow, for that given data assessment critieria. For example if the judgement is that, for a given parameter, the data needs to be inflated by 20% to match true flow (there is undercount) then the median value on the linear scale needs to be set at 1.2 on the linear scale, as the effect is multiplicative, where true flow = adjustment factor × the data. This is then transformed to the log-scale to provide the θ parameter.

The value of each of the τ hyperparameters is elicited from a judgement of either the 2.5 percentile or the 97.5 percentile of the distortion of true flow. Going back to the illustrative example above, if from an available source of information or the data assessment carried out in chapter 3 the judgment is that there a 1/40 chance that this undercount could be as higher than 50%, then in this case q_h is elicited as 1.5 on the linear

scale. To calculate the standard deviation on the log-scale, and thus the precision hyperparameter, equation 6.8 is applied to the elicited median and 97.5 percentile. This process is repeated for each of the priors that are assumed to take a normal distribution, and the results are reported in table 6.2.

As mentioned in chapter 5.4.2, the 'true flow' terms in both the student and non-student data assessment equations act as balancing terms in the model. In effect all of the judged distortion of the data assessment parameters is balanced out in the z_{it}^{S} and the z_{it}^{W} terms. As specified in chapter 5.4.2, both these parameters are assumed to take a log-normal distribution. Subscript *i* denotes the country of citizenship and subscript *t* denotes the year. As such there are 279 posterior distributions estimated for both z_{it}^{S} and z_{it}^{W}, but only a single prior is specified for a term. Seeing as though the role of the true flow parameter is to act as a balancing term in the model, and it is where we get the estimation of citizenship specific immigration from, the prior needs to be elicited in way that takes into account the existing range of estimates of UK immigration. The median and precision hyperparameters are elicited using quantiles and equations (6.8) and (6.9) are used to calibrate the precision term τ.

Most of the priors in the student and non-student Data Assessment Models are assumed to be normally distributed. However, the priors for the def^{HESA}, def^{DWP} and $def^{DWP.EU}$ are specified as taking a gamma distribution, $\Gamma(r, \alpha)$, where r is the shape parameter and α the scale parameter. Under the assumed parameterisation of OpenBugs (Spiegelhalter et al 2011), the mean of the gamma distribution is $\frac{r}{\alpha}$ and the variance of a gamma distribution is the shape parameter divided by the square of the scale parameter: $\frac{r}{\alpha^2}$. Consequently, values for the mean, θ, and variance, σ^2, need to be elicited so that the shape parameter can be calculated through a method of moments,

$$r = \frac{\theta^2}{\sigma^2} \tag{6.10}$$

and the scale parameter is calculated by equation 6.11:

$$\alpha = \frac{\theta}{\sigma^2}. \tag{6.11}$$

The mean is relatively straight-forward to elicit, as it is simply a quantification of the distortion of true flow. For example, if a data set does not have complete coverage as it

does not include children, then the data assessment parameter needs to inflate true flow. So, on the natural scale, θ needs to be greater than 1 as it is interpreted as a multiplicative adjustment, where true flow = adjustment factor × data. So a θ value of 1.2 would inflate true flow, relative to the data, by 20%.

Eliciting a value for the variance is less straight forward. The quantile method proposed earlier in this section, is based on assumptions of normality. Unlike a normal distribution, the gamma distribution is positively skewed. Its skewness is equal to $2/\sqrt{r}$; and only depends on the shape parameter, r. However, as the shape parameter increases a gamma distribution becomes more like a normal distribution and is more symmetric (Patel and Read 1996). For example if x follows a random gamma distribution $\Gamma(r, \alpha)$ and y is a normal random variable with the same mean and variance then $F_x \approx F_y$ when the shape parameter is large relative to the scale parameter and where F denotes the probability distribution function.

For each of the three parameters that take a gamma distribution the prior, based on the calculation of the shape and scale parameters, is fitted in Open Bugs where a detailed comparison is made between the elicited and the fitted quantiles. In general the differences between the quantiles for the elicited and fitted distributions are only small. This is a particular limitation with regard to this method of prior elicitation, and further research could attempt to take this skewness into account during elicitation of subjective judgement.

A gamma distribution is still an appropriate choice for these particular definition parameters, though. The priors need to be strictly positive and we have more certainty closer to one on the linear scale and zero on natural scale, making a positively skewed distribution an appropriate expression of our prior beliefs of the distortion of true flow.

6.4.2 Elicitation of Hyperparameters

The contribution of this section is a detailed documentation of how each of the hyperparameters is elicited and the subjective judgements they are based on. This is important for the Data Assessment Model, as the prior values are of significant consequence in the estimation of the true flow posteriors because the model is not constrained to a given standard statistical framework, such as the Bayesian log-linear.

Where there is useable evidence of a distortion of true flow, as summarised in table 3.1 in chapter 3, this is used in the elicitation.

Model	Parameter	Distribution	Median/Shape	Precision/Scale
Student Data Assessment Model	$def^{IPS.S}$	Normal	$\theta_{def}^{IPS.S} = 0$	$\tau_{def}^{IPS.S} = 3246$
	$cov^{IPS.S}$	Normal	$\theta_{cov}^{IPS.S} = 0$	$\tau_{cov}^{IPS.S} = 3246$
	$cov^{IPS.EU.S}$	Normal	$\theta_{cov}^{IPS.EU.S} = 0.18$	$\tau_{cov}^{IPS.EU.S} = 115$
	$bias^{IPS.S}$	Normal	$\theta_{bias}^{IPS.S} = 0$	$\tau_{bias}^{IPS.S} = 1613$
	def^{HESA}	Gamma	$r_{def}^{HESA} = 18$	$\alpha_{def}^{HESA} = 185$
	cov^{HESA}	Normal	$\theta_{cov}^{HESA} = 0.38$	$\tau_{cov}^{HESA} = 2266$
	z_{it}^{S}	Log-Normal	$\theta_{z}^{S} = 8$	$\tau_{z}^{S} = 0.2$
Non-Student Data Assessment Model	$def^{IPS.W}$	Normal	$\theta_{def}^{IPS.W} = 0$	$\tau_{def}^{IPS.W} = 3246$
	$cov^{IPS.W}$	Normal	$\theta_{cov}^{IPS.W} = 0$	$\tau_{cov}^{IPS.W} = 1613$
	$cov^{IPS.EU.W}$	Normal	$\theta_{cov}^{IPS.EU.S} = 0.34$	$\tau_{cov}^{IPS.EU.S} = 34$
	$bias^{IPS.W}$	Normal	$\theta_{bias}^{IPS.W} = 0$	$\tau_{bias}^{IPS.W} = 1613$
	def^{DWP}	Gamma	$r_{def}^{DWP} = 3.84$	$\alpha_{def}^{DWP} = 79$
	$def^{DWP.EU}$	Gamma	$r_{def}^{DWP.EU} = 19$	$\alpha_{def}^{DWP.EU} = 108$
	cov^{DWP}	Normal	$\theta_{cov}^{DWP} = 0.095$	$\tau_{cov}^{DWP} = 38$
	z_{it}^{W}	Log-Normal	$\theta_{z}^{W} = 8$	$\tau_{z}^{W} = 0.2$

Table 6.2: Details of the Distribution and Elicited Values of Each Prior for the Data Assessment Model

In this section the second and third stages of prior elicitation proposed by Garthwaite et al (2005) are carried out. As mentioned previously, these stages often happen concurrently, which also is the case for the priors of the Data Assessment Model. Elicitation is carried out for each of the priors specified in chapter 5.4.2. The parameter of interest, probability distribution specified, and the values elicited for each of the hyperparameters is detailed in table 6.2 below. Note that each of the hyperparameters are detailed on the log-scale in the table. For ease of interpretation, the elicitation throughout this section is mainly discussed on the linear scale.

The first priors to be elicited are for the true flow parameters z_{it}^{S} and z_{it}^{W}. As outlined in the previous section, the role of the true flow priors is to provide a reasonable range for all the estimates of true flow in both the student and non-student models. As such they are centred at 8 on the log scale and given a precision of 0.2. This leads to a 2.5 percentile value of 36 on the linear scale and a 97.5 percentile value of 234,100 on the linear scale. This credible interval should cover the range of estimated values of the true flow priors.

Where it is difficult to elicit the magnitude of the distortion of true for a given data assessment parameter, but there is good reason to believe that there is a distortion of true, then the prior is centred at one on the linear-scale and a judgement of uncertainty is elicited using the quantile method outlined earlier in section 6.4. This is the case for the bias parameters for the IPS for both the student and non-student model. The IPS is an intentions-based survey, and the bias parameters are designed to take into account respondents whose intention changes after interview. Data on migrant and visitor switchers, the distortion detailed in the data assessment of chapter 3, was not made available for this research. However, as outlined in chapter 5.4.2, we do know that the distortion undercounts immigration in the case of migrant switchers, and overcounts immigration in the case of visitor switchers. As such, with the median centred at 1 on the linear scale, and the assumption of normality, the $bias^{IPS.S}$ and $bias^{IPS.W}$ priors allow for both the effect of visitor and migrant switchers to be taken into account.

With regard to the effect on the estimate of true flow posterior, eliciting the bias priors in this way has the effect of propagating uncertainty in the estimate of UK immigration. As such, in setting the $\tau_{bias}^{IPS.S}$ and $\tau_{bias}^{IPS.W}$ hyperparameters the main consideration is making a judgement about the level of uncertainty we think migrant and visitor switchers should be propagated in the true flow term. For both students and non-students, the 97.5 percentile was elicited to be 1.05, which means that the judgement is that there is a 1/40 chance that there could be an undercount higher than 5% as a result of the bias assessment criteria. Equation 6.9 is applied to set the precision hyperparameters for students and non-student, which are detailed in table 6.2.

Where the data is deemed a good reflection of true flow, then the prior is centred at 1 on the linear scale. This is the case for the $def^{IPS.S}$, $def^{IPS.W}$, $cov^{IPS.S}$ and $cov^{IPS.W}$ (non-EU) parameters. With regard to definition, the IPS is a good reflection of true flow, so a high level of precision is set through eliciting the 97.5 percentile at 1.035. There is

more confidence in the definition data assessment for the IPS than the bias assessment; hence, the elicitation of a more certain prior. This results in a precision value that is approximately twice the level elicited for the bias priors (see table 6.2).

The prior for the coverage parameter $cov^{IPS.S}$ for non-EU student immigration is elicited in a similar way. Referring back to chapter 3.3.1 we know that there is a high level of survey coverage at the main hub-airports. A large proportion of non-EU immigration passes through Gatwick and Heathrow. Furthermore, Chinese and Indian citizens make up a large proportion of the non-EU student flow and they are highly likely to travelling into the UK via Heathrow. So, the judgement is to centre the prior on 1 on the linear scale and to set the precision, having elicited the 97.5 percentile to be 1.035. For $cov^{IPS.W}$ the precision was calculated from an elicitation of 1.5 for the 97.5 percentile. This leads to a slightly higher level of uncertainty non-student flows from outside the EU. This is because this flow is less dominated by Indian and Chinese migrants, and therefore, it is possible ports and airports that are surveyed less are used by these people.

The IPS coverage parameter for EU students and non-students needs to inflate true flow, as there is a possible undercount of migrants who travel into the UK via regional airports and ports (ONS 2012 c). The effect of this undercount is judged to be higher for non-students, due to the high levels of labour migration from A8 countries following the expansion of freedom of labour movement after 2004. The ONS have estimated that this undercount was substantial, an average of 40,000 per year from 2005-2010 (ibid, page 16). This is approximately 40% of the total inflow of EU non-student migrants, as estimated by the IPS. As such, the median for $cov^{IPS.EU.W}$ is set at 1.4 on the linear scale. This is the only source of information on the level of undercount that can be attributed to this parameter. However, we can say with a reasonable level of certainty that the IPS does underestimate immigration for EU non-students. As such, the 2,5 percentile is elicited with caution and is set at 1. This results in a relatively high level of uncertainty for this prior. There is also a lower level of inflow of students from Europe. As such, the median is set at 1.2 for $cov^{IPS.EU.S}$

The following parameters are elicited using specific evidence and from previous studies. For the definition parameters of both sources of administrative data, the prior needs to deflate true flow as there is a definite overcount of immigration. For the DWP data, this is a result of short term migrants registering for NINo and for the HESA data this is a result of students who either do not complete their first year of study.

In section 5.4.2 a gamma distribution is assumed for def^{HESA}, def^{DWP} and $def^{DWP.EU}$ as the distribution, to be a realistic representation of the distortion caused to true flow, needs to be strictly positive. As outlined in section 6.4.1 there is a limitation to the way the uncertainty is elicited for gamma distributed data assessment parameters, a comparison of the fitted quantiles to the ones used in elicitation is detailed in table 6.3.

For def^{HESA} the median value is elicited from data on the drop-out rate of for all students, including British citizens. The proportion of all students who continue onto their second year of study ranges from 87.8% in 2002/03 to 90.7% in 2010/11 (HESA 2012). The majority of migrants who do not continue on to their second year will probably not have been resident in the UK for a period of at least 12 months, hence distorting true flow. The drop-out rate for all students is assumed to be approximately the same for non-UK domiciled students, consequently the median is set at 1.1. Due to the nature of the model specification in chapter 5.4.2 this should have a deflationary effect on true flow. The variance is calculated by eliciting the 97.5 percentile as 1.2. The shape and scale hyperparameters are calculated by applying equations 6.10 and 6.11.

The DWP definition parameters are split into an EU group and a non-EU group. Following the expansion of the freedom of labour movement in the EU, there was a substantial amount of labour migration to the UK. For the EU group, there is probably a large number of short term migrants who register for a NINo, especially following the well documented expansion of freedom of movement in the EU. The prior median for this parameter is taken from the expert elicitation using a Delphi survey method, for the IMEM model (Raymer et al 2013). Details of the IMEM model specification are outlined in chapter 5.3.2 and the prior elicitation in section 6.2.2. The experts consulted in the IMEM elicitation judge that for a six month definition there is a 19% percent overcount in relation to true flow, so the mean for $def^{DWP.EU}$ is set at 1.19. With the true flow acting as a balancing term in the model, this has a deflationary effect on true flow in relation to the data. The variance was elicited from the 2.5 percentile being set at 1.1, as we know from the assessment of data in chapter 3 with a high level of certainty that there is an overcount in the DWP data.

For the non-EU group, there is still probably overcount in the DWP data. However, as there is probably fewer short term migrants from outside the EU, the mean for $def^{DWP.EU}$ is set at 1.05 with the variance elicited from the 2.5 percentile being set at 1.

As outlined in chapter 3, the HESA data only covers students at publicly funded HE institutions, with language school and FE students not being covered. Consequently, there is an undercount in the HESA data. The Home Office have data on visa applications for FE colleges, language schools, independent schools and an 'other' category (Home Office 2012). In 2010 there was a total of 106,037 applicants for study visas for non-EU students. The median is then set by calculating the adjustment factor required to inflate the HESA total of non-EU students for 2010 by 106,000. This provides a median value of 1.47 on the linear scale. It is assumed that the level of undercount for EU students is the same. The uncertainty for cov^{HESA} was elicited by setting the 97.5 percentile at 120,000 and performing the same calculation to determine the adjustment factor.

The DWP coverage parameter, due to the assessment carried out in chapter 3, needs to reflect that there is under coverage of children and people who either do not claim any kind of social security or do not work, but there is over coverage of students who work part-time. There is no strong evidence about the level of distortion this has on true flow, so this prior is elicited from subjective judgement alone. As such, the median for cov^{DWP} is elicted at 1.1 on the linear scale and the 2.5 percentile is set at 0.8. The effect of changing the prior assumptions for cov^{DWP} is checked in the sensitivity analysis of chapter 7.

6.4.3 Discussion and Limitations

As outlined in chapter 5 and in section 6.3 of this chapter, the prior elicitation for the Data Assessment Model is of consequence in the estimation of true flow. The model is not marginally constrained to fit given values, unlike the Bayesian Log-linear model. Consequently, the hyperparameters need to be elicited with care, and based on the best evidence available. This is outlined in detail in section 6.4.2, where the subjective judgements for each of the parameters of the Data Assessment Model is made.

The aim of the elicitation is to capture the assessment of each of the data sources carried out in chapter 3, determining the judged nature and uncertainty of the distortion of true flow for each of the sources of data. As such, the method of eliciting quantiles that was outlined in section 6.3.3 is applied for all the parameters. Through this elicitation process, the gathering of evidence from chapter 3 and other sources where possible in the previous section, the summary table 3.2 of the data assessment can now effectively be

modelled explicitly in an estimation of UK immigration. This is the main contribution of the elicitation carried out for the Data Assessment Model.

The elicitation is not without its limitations, however. The elicited gamma distributed priors do not quite match their fitted values. This is a result of the quantile method of elicitation being conducted under assumptions of normality. Furthermore, it is difficult to consider every single distortion of true flow in the modelling framework as more data assessment parameters would need to be specified. The Data Assessment Model, does however, provide a framework for the main thrust of the distortions of true flow to be captured via the priors. Further research could use the elicitation of subjective judgement from a panel of experts, using a Delphi approach similar to that of Bijak and Wiśniowski (2010) and Wiśniowski, Bijak and Shang (2014). These approaches are detailed in section 6.2.2.

6.5 Conclusions

The methods of elicitation proposed in this chapter are an important contribution to the continued development of modelling statistics, where subjective judgements of the data available are modelled explicitly, through parameters in the model. Simply identifying the evidence, which provides an assessment of the quality of immigration data alone, is not an adequate contribution. In the case of this research, only when this evidence is effectively translated into prior distributions, which are a realistic representation of the distortion of true flow of immigration, according to the available evidence on data collection, has a significant contribution been made.

The aim of the elicitation in this chapter has been to come up with prior distributions which bridge the gap between what we know about each source of data and the parameter in the model for which it is set. The two different modelling approaches outlined in chapter 5 provide the framework for the data assessment of chapter 3 to be modelled explicitly. A further contribution of this chapter is bridging the gap between the qualitative assessment of the data in chapter 3 and the model parameters specified in chapter 5, through the use of the best available evidence. For the Bayesian Log-linear model, an approach to eliciting prior means (equivalent to median for a normally distributed prior) has been suggested. In the literature, there are very few examples of elicitation of informative priors for a Bayesian Log-linear model. Bridging the gap between

the prior and the main effects in the model in section 6.3.1 is a small contribution to furthering understanding of how this can be carried out in an applied setting.

A method using quantiles, is proposed to overcome the problems of eliciting a value for the precision on the log-scale. Using the quantile approach outlined in section 6.3.2 and applied in that section and section 6.4.2 allows a credible interval to be elicited on the linear scale, and then the uncertainty to be calculated on the log scale from this. Through eliciting priors for both the Bayesian Log-linear model and the Data Assessment Model, it is clear that prior elicitation is more straight-forward where the model in question is closely related to an assessment of true migration flow.

Chapter 7 – Results of Bayesian Models

7.1 Introduction

Following chapters 5 and 6, full probability models have been specified and informative priors have been elicited for the Bayesian Log-linear model and the Data Assessment Model. Chapter 5 provided the framework for the subjective informative elicited in chapter 6 to be explicitly included in two different statistical models of immigration. The aim of this chapter is the computation and sensitivity analysis of both models and a reflection on the results.

As introduced in chapter 6.2.1, it is recognised that elicited prior distributions are both not perfect representations of subjective judgements and are changeable dependent on available evidence and the method of elicitation. To test how this affects the estimate of the posteriors there is a widespread use of sensitivity analysis in Bayesian applications (O'Hagan and Forster 2004). The most common use of sensitivity analysis has been to vary a small number of the prior probability distributions in an ad hoc way (Garthwaite 2005).

In models where multiple priors are specified there are a large number of combinations of possible sensitivity analysis. It is unrealistic, and arguably, not possible to carry out a fully comprehensive sensitivity analysis in the strictest sense. As a result, in this chapter, the sensitivity of the Bayesian Log-linear model and the Data Assessment Model to a selected number of different prior assumptions is computed. This is organised in a way to be recommendation focussed. Through a comparison of the posteriors computed, from the models specified with informative priors in chapter 6 and the posteriors from the sensitivity analysis, one can make substantive conclusions about the distortion of true flow, within the framework of the data assessment criteria outlined in chapter 3.

Firstly, a brief outline and introduction of Bayesian model computation is given in section 7.2. In this section the various diagnostic tests, that are used to check model convergence, are outlined. Following this there is an account of the computation of the Bayesian Log-linear model in section 7.3. Firstly there is a consideration of a model fitted with vague priors specified for each of the log-linear parameters. This provides posterior estimates where the data, given the log-linear framework, are dominant. Then the results of the four set of model estimates with informative priors are compared to the posteriors

estimated with vague priors. For the Bayesian Log-linear model, the prior elicitation for the sensitivity analysis was carried out in chapter 6.3.2.

Following this, the Data Assessment Model is computed and the results are presented in section 7.4. Through comparisons of the posterior estimates to the data used in the model, initial conclusions are made about the sensitivity of the estimates of UK immigration to the distortion of true flow, propagated from the prior elicitation of the data assessment criteria.

A brief comparison with the Bayesian log-linear estimates is made and a sensitivity analysis is conducted. Examples of a combination of different changes in prior judgements are made in order to test the sensitivity of the posterior estimates to varying the assessment of true flow. As such, in the reporting of results, and to make full use of taking a Bayesian approach, it is important to consider summary results about the whole posterior distribution. Simply reporting just point estimates would suppress all the information propagated into the model from the priors and would not help aid our understanding of the uncertainty inherent in UK immigration statistics.

7.2 Bayesian MCMC Computation Using OpenBugs

As outlined in chapter 5.2.2, having specified a full probability model, the second step in an applied Bayesian analysis is to *update the knowledge specified in the full probability model about the unknown parameters by conditioning on the available observed data* (Gelman et al 2009). Effectively this means computing the model to produce estimates of posterior probability densities. These posteriors include all the information propagated in the model from the data and the priors elicited in chapter 6, providing an expression of uncertainty for each of the parameters, as specified in chapter 5. As it may not be possible to estimate the posteriors analytically, a numerical method is required. Consequently, Markov Chain Monte Carlo (MCMC) methods are used to estimate the posteriors of each of the stochastic parameters in the model.

In section 7.2.1 there is a brief introduction to MCMC computation as used within the OpenBugs software (Spiegelheiter et al 2011). Care needs to be taken when applying MCMC methods. MCMC is an iterative solution to model estimation, and there are generally two issues. The first is determining if the underlying numerical algorithm based

on a Markov Chain has reached the stationary state, which is the desired posterior distribution. This is known as convergence. Secondly, one needs to determine how many iterations to keep after convergence to ensure an accurate estimate of the posterior. There is no single test for convergence; instead there are numerous diagnostics one can compute. The tests of convergence applied later in this chapter follow the brief introduction to MCMC methods.

Having specified the model as a full joint distribution on all quantities, whether parameters or observables, to estimate the posterior distribution we need to sample values of the unknown parameters from their conditional distribution given the data and the priors (Spiegelhalter et al 2011). Applied Bayesian methods have become closely linked to sampling-based estimation methods because models are often too complex for direct computations (Congdon 2006). In general Markov chain Monte Carlo methods (MCMC) are used and the software utilised in this research to carry out this task is OpenBugs (Spiegelhalter et al 2011).

The long-run stationary distribution of the MCMC algorithm is the estimate of the posterior distribution. It is beyond the scope of this thesis to go into great detail about MCMC. However, through using OpenBugs one does experience some of the key features of MCMC such as the 'burn in' period required and the need to produce numerous simulations to estimate posterior distributions accurately. It is therefore necessary to outline some its characteristics and the need for visual and numerical assessment of convergence of the underlying Markov chains.

Through running one or more chains we get a very large number of simulated values, the distribution of which approximates to the posterior distribution. To help check convergence it is useful to run more than one chain. There is no single comprehensive check of model convergence. As such, various tests of convergence are outlined in this section, which are then applied for each model estimate in sections 7.3 and 7.4 later in the chapter. When checking model convergence it is necessary to monitor all of the parameters in the model, rather than just the parameters of interest. Convergence may be apparent and be achieved after relatively few iterations for certain stochastic nodes, whereas for others it may never be reached, or requires a long run of the algorithm.

Convergence simply means the point at which it is reasonable to assume the samples generated from the simulation are representative of the underlying long-run

stationary distribution of the Markov chain (Gelman et al 2009). In practice, there is a certain amount of ignorance about how quickly convergence has occurred, and we have to fall back on post hoc testing of the sampled output (Plummer et al 2006). By convention, the samples produced before convergence is reached are discarded, thus splitting the samples into two sections. The section of the computation including samples where convergence has yet to be reached is known as the 'burn in' period. When convergence is reach the burn in iterations are discarded. Further iterations are then computed to provide an accurate estimate of the posterior.

The first way of monitoring convergence is a visual analysis of the trace and history plots in OpenBugs (Spiegelhalter et al 2011). If the plots from each respective chain overlap one another then we can be reasonably confident that the model has converged. This will be the primary way that convergence will be assessed in this research.

The second way of monitoring convergence that is applied is the Brooks-Gelman-Rubin (BGR) statistic, as modified by Brooks and Gelman (1998). The BGR statistic provides an assessment of the stability both within and between Markov chains (Congdon 2006). It is represented as a plot where the width of the pooled and average BGR statistic should be stable and their ratio should be equal to one (Spiegelhalter et al 2011).

Parameter samples obtained by MCMC may be heavily correlated, which means that extra samples are needed to convey the same information (Congdon 2006). One can obtain a plot of autocorrelation from the OpenBugs software, where if autocorrelation does not vanish to zero, less information about the posterior distribution is being gained from each distribution. This is the third model diagnostic that will be obtained from a plot. The fourth and final diagnostic plot used is a 'running mean' and 'running quantiles', which can be also plotted and checked for stability (Bijak 2010). The quantiles calculated here are the 95% confidence intervals of the mean, which is calculated at each iteration using the whole backward sample.

After establishing convergence, one needs to run further iterations to ensure that the posterior estimates are accurate. One way to assess this is to calculate the Monte Carlo Error for each parameter, which is reported as a part of standard output in OpenBugs. This is an estimate of the difference between the mean of the sampled values, which is used as the estimate of the posterior mean, and the true posterior mean (Spiegelhalter et al 2011). A generally acceptable Monte Carlo error for each parameter of interest is less than

5% of the sample standard deviation (Spiegelhalter et al 2011). To ensure that the Monte Carlo Error is small in estimating the posterior and other points of interest (such as the mean, median and other quantiles on the posterior), enough observations need to be taken after the initial burn in period.

7.3 Results of the Computation for the Bayesian Log-linear Model

Having specified the full probability model in chapter 5.4.1 and elicited informative priors in chapter 6.3, the Bayesian Log-linear model has been estimated. Convergence is checked for each of the model estimates using the diagnostic plots detailed previously, and the accuracy of the posteriors is checked by computing the Monte Carlo error for each parameter. To assess the sensitivity of the data, given the log-linear likelihood, to informative priors a model with vague prior assumptions is estimated in section 7.3.1. This then provides a benchmark against which posterior sensitivity to prior assumptions can be checked. The results from the four informative prior models are presented in section 7.3.2. The comparison of results from the posterior estimates under vague and informative priors provides the basis for the substantive conclusions of section 7.3.3.

7.3.1 Posterior Estimates Under Vague Priors

Vague priors are specified when, for a given statistical model, the posterior is required to describe whatever the data 'have to say' about a particular quantity of interest (Bernardo 1996). Throughout the analysis proper prior distributions that integrate to one are used. However, where the posterior is required to reflect the data, prior distributions are given very large variances.

This produces posterior distributions which are driven by the data, and as such do not include any additional information about data assessment. Here, vaguely informative priors are used as technical devices to produce posteriors, which are a reflection of what the data is telling us about UK immigration, via the formal use of Bayes' theorem. In effect the role of a vaguely informative prior, in a given model and for particular parameter within this model, is to make the data dominant (Bernardo 2003).

For the Bayesian Log-linear model, the classical analysis of a GLM is obtained if a flat, non-informative distribution is specified for the parameters in the model (cf Gelman et al 2004). In this case the estimate of the posterior distribution is close to the maximum likelihood estimate for the parameters in the model, with posterior inference coming from a normal approximation to the likelihood (ibid).

With the above in mind, the priors for the Bayesian Log-linear model are assumed to be normally distributed with a median of zero on the log-scale and a precision of 0.0001. This produces priors which are extremely diffuse. The posterior characteristics are computed with an initial burn in of 5000 iterations, which were discarded as convergence had been reached. A further 25,000 iterations were computed with a thinning of 10 (ie each 10^{th} iteration was taken). The chain is thinned to reduce autocorrelation, whilst allowing a long run, without having to save each iteration. The Monte Carlo error is checked, and is deemed acceptable as it is estimated to be less than 5% of the standard deviation for each parameter in the model.

A frequentist log-linear model, for both students and non-students is estimated for the top 10 flows and as expected, the results of the Bayesian Log-linear model, under vague priors reflect that of a classic GLM. The true flow posteriors have a very high level of certainty. This is a result of not having the survey weights, and is the same problem that is encountered in chapter 4 with the Bayesian log-linear model.

The posterior estimate of true flow has a credible interval which is unrealistic. Given the disparity in the two sources of data, HESA and IPS, and the assessment of the data in chapter 3, the credible intervals should indicate that there is more uncertainty in the posterior estimate of true flow. This is a limitation of the survey weights for the IPS not being made publicly available.

7.3.2 Posterior Estimates Under Informative Priors

In chapter 6.3.2 priors have been elicited for four separate models, to assess the sensitivity of the model to varying assessments of the auxiliary marginal priors elicited from the Census and Flag 4 data respectively. The values for the prior hyperparameters are detailed in table 6.1.

Firstly the model is computed under priors which, for a given main effect in the model, have the same value for all levels of *i* and all levels of *t*. For the first model-estimate the posterior characteristics are computed with an initial burn-in of 15000 iterations. A further 25,000 iterations were computed with a thinning of 10 (ie each 10th iteration was taken). Initially the model was estimated with the following precision values: $\tau^{[1]}_{\lambda_i^S}$, $\tau^{[1]}_{\lambda_t^S}$, $\tau^{[1]}_{\lambda_i^W}$ and $\tau^{[1]}_{\lambda_t^W}$ (see table 6.1 for details). The superscript for each of the precision terms in this section refers to the level of precision elicited for a given computation. The non-student posteriors are insensitive to this initial prior specification and the results are very similar to the posterior computation under vague priors.

However, there is sensitivity to the priors in the student model. Firstly, it is necessary to compare the posteriors of the overall effects, as these are the benchmarks each of the main effects in the model is set in relation to. Any difference in the posterior value for λ^Shas not been driven by the main effect prior as this is assumed to be vague. When changing a prior for a given parameter has an effect on a different parameter in the model this is referred to from this point on as a 'secondary effect'. A secondary effect on the overall effects, where there is sensitivity in the posterior main effects, is to be expected, to satisfy the constraint of the model to log-linear margins.

Under vague priors for all parameters the posterior median and interquartile range (in brackets) of $\lambda^S$is 1.114 (1.113, 1.116) and under precision level [1] they are 1.490 (1.492, 1.495), all on the linear scale. The posterior characteristics are quoted to three decimal places as there is a large amount of certainty in the estimates. This increase, however, is balanced out by a proportionate decrease for all posterior estimates of λ_i^S. For example, for Australia the posterior median and interquartile range for λ_i^S under vague priors is 0.825 (0.832, 0.840) which then decreases under precision level [1] priors to 0.587 (0.582, 0.592). In effect, even though the origin specific λ_i^S posteriors seem sensitive to the priors, there is no relative change in the marginal effects. This result is as one would expect, as there is a low level of certainty expressed in the priors for the country-specific main effect.

For the λ_t^S posteriors, however, there is a proportionately larger increase for 2002 and 2003 and proportionately smaller increase for 2009. For example, the λ_t^S posterior estimates of the median and interquartile range for 2002 under vague priors is 0.998 (0.996, 1.001) which increases to 1.332 (1.329, 1.335) under precision level [1] priors.

This results in student estimates of true flow, where for the first few years (especially 2002 and 2003) of the study period there is an increased estimate of immigration. This is illustrated by figure 7.1 below which plots the posterior median of true flow under vague and precision level [1] priors for Pakistani students. Due to the log-linear constraints the increase in 2002 and 2003 is balanced out by a decrease in the final years of the study.

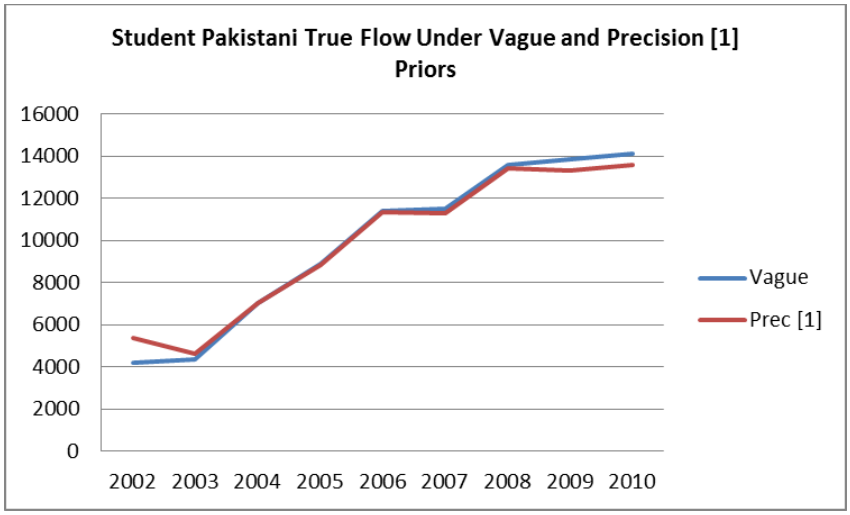

Figure 7.1: Comparison of Posterior Medians for Pakistani Student True Flow under Vague and Precision [1] Level Priors

The posterior credible interval (interquartile ranges) cannot be plotted as they would not be visible, because there is an unrealistic level of certainty. In 2002 there is a 20% level of sensitivity to the prior assumptions and the quantity of this increase is balanced out in 2009 and 2010. Effectively, for students, the sensitivity of the Flag 4 priors indicates that there is a small undercount of students in 2002 and 2003 according to the log-linear model.

To investigate the level of sensitivity of the log-linear estimates when confronted with the Flag 4 auxiliary data as more certain quasi-alternative marginal effects the Flag 4 prior precision is increased by an order of magnitude for both students and non-students. The model is then computed, with the same number of iterations as above under priors with the following precision values: $\tau^{[2]}_{\lambda_i^S}$, $\tau^{[2]}_{\lambda_t^S}$, $\tau^{[2]}_{\lambda_i^W}$ and $\tau^{[2]}_{\lambda_t^W}$. Here we are increasing the level of confidence we have in the Flag 4 data as an alternative to the IPS for λ_i^S and λ_i^W.

The model is computed in the same way as for precision level [1] with convergence reached after 15000 iterations. Again the non-student main effects in the model are, on the whole, insensitive to the prior assumptions. There is a very small amount of sensitivity in the posterior estimates of λ_t^W. Whether this increases with even greater levels of certainty expressed in the priors is tested in the final sensitivity analysis when the precision level [4] priors are used in the computation. For students, the posterior estimates are sensitive to almost exactly the same degree, even with increased certainty expressed in the priors.

With the lack of sensitivity in the origin specific posteriors, the next stage is computed the model with separate priors elicited for each country. As outlined in chapter 6.3.2. The posteriors are then computed with country-specific priors set for non-students and because the Census data is not judged to be a good reflection of a true flow marginal effect for students this prior is kept relatively vague. The Flag 4 priors are kept the same as for the second set of precisions.

For this model estimate convergence is more straight forward and an initial burn in period of just 10,000 iterations is required. The precision levels, detailed in table 6.2 in the previous chapter, are as follows: $\tau_{\lambda_i^S}^{[3]}$, $\tau_{\lambda_t^S}^{[3]}$, $\tau_{\lambda_i^W}^{[3]}$ and $\tau_{\lambda_t^W}^{[3]}$. For the student and non-student origin specific parameters the posteriors are insensitive to the country-specific priors and there are no secondary effects.

Finally the precision levels for all parameters are tightened by a further order of magnitude with precision levels of $\tau_{\lambda_i^S}^{[4]}$, $\tau_{\lambda_t^S}^{[4]}$, $\tau_{\lambda_i^W}^{[4]}$ and $\tau_{\lambda_t^W}^{[4]}$. The priors used in the computation now have an unrealistically high level of certainty, in comparison to the judgements made in chapter 5. However, the aim of conducting the sensitivity analysis in this way is to determine the level of confidence we would need to have in the Flag 4 and Census data, respectively, as auxiliary marginal effects in a log-linear estimate of true flow.

There is no sensitivity in the posterior estimates of tightening the origin specific priors. This suggests that in the estimate of the margins of origin-specific true flow, the data is dominant in comparison to the Census. This corroborates with the detailed assessment of the Census data carried out in the prior elicitation where a high level of uncertainty was elicited for the marginal effect priors.

With regard to the non-student year specific parameter λ_t^W the posterior estimates are only slightly sensitive to the precision level [4] priors. Proportionately, however, the non-student overall effect is slightly more sensitive to the priors with a posterior median and interquartile range of 0.227 (0.226, 0.228) in comparison to 0.212 (0.213, 0.231) under vague priors. With the main effects largely unchanged, this secondary effect leads to estimates of true flow for non-students where the posterior median for 2010 is slightly lower. Figure 7.2 below illustrates this slight sensitivity.

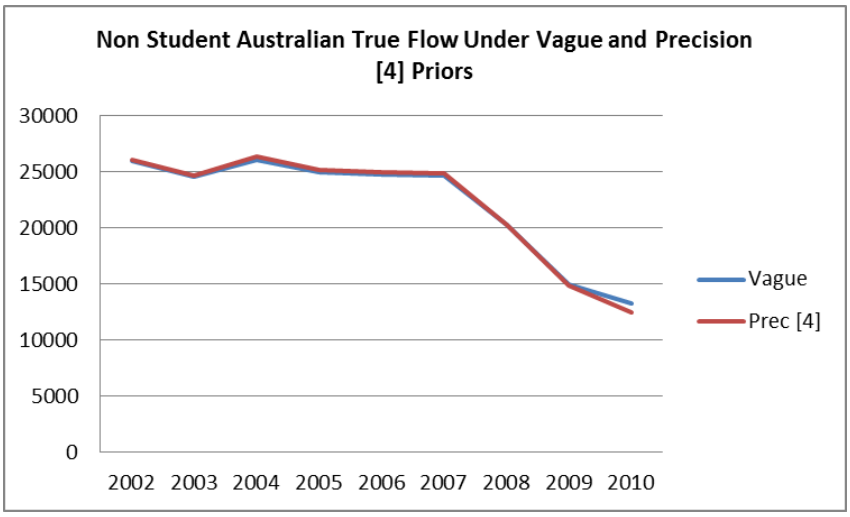

Figure 7.2: Comparison of Posterior Medians for Australian Non-Student True Flow under Vague and Precision [4] Level Priors

Finally, there is only a slight increase in the sensitivity of the posterior student parameter estimates in comparison to under the previous levels of precision. As result, for students, it seems that marginal time-specific true flow is equally sensitive to the Flag 4 prior being set with a low and high precision.

,

7.3.3 Discussion and Limitations

With the model being constrained to log-linear margins, testing the sensitivity to prior assumptions has to be carried out in relation to the overall effects in the model. If the sensitivity of the posteriors simply mirrors the sensitivity in the overall effects then the estimate of true flow is unchanged. This is the case for students throughout each of the four model estimates. This could be as a result of the priors for the overall effect remaining

vague. As outlined in chapter 6, under the current model corner cell constraints, specifying informative priors for the main effects is problematic.

With regard to practical recommendations that can be made from this results of the model, it is clear that the student year-specific margins are slightly more sensitive to the Flag 4 data than for non-students. However, overall the results are not sensitive the alternative marginal effects. As such, within the constraints of the log-linear framework, the IPS margins remain dominant when confronted with highly certain alternative data.

This is the only part of true flow – our trust in the IPS margins – that can be tested in this model. To take into account possible bias in the IPS and problems with undercount due to lack of coverage at regional airports for example an unconstrained Bayesian model is required. The next stage in of this chapter is to investigate the results of the Data Assessment Model, which attempts to explicitly model these (and other) considerations of the distortion of true flow.

7.4 Results of the Computation for the Data Assessment Model

Given the specification of the model in chapter 5.4.2 and the elicitation of the data assessment priors in chapter 6.4.2, the full probability model is computed in this section. Convergence and the Monte Carlo Error for each of the parameters are checked in the same way as section 7.3. Firstly, posteriors are computed under the informative priors elicited in chapter 6.4.2. An initial assessment of posterior sensitivity to the priors is carried out by comparing the computed parameter estimates to the data in section 7.4.1. These estimates are also briefly compared to the results of the Bayesian Log-linear model detailed in the previous section.

With the Data Assessment Model being unconstrained, sensitivity analysis is important. This is carried out in section 7.4.2. Examples are used to illustrate how changing the subjective judgement of the data assessment criteria affects the posterior estimates of true flow.

7.4.1 Posterior Estimates Under Informative Priors

Posterior estimates of the Data Assessment Model are computed with the priors taking informative probability distributions, which were elicited in chapter 6.4.2. Following a burn in period of 7000 (thinned for every 100th) iterations each parameter converged. Each of the convergence tests outlined in section 7.2 were checked to determine this. A further 8000 iterations were computed, also with a thinning of 100, to ensure an accurate estimate of the posterior. Monte Carlo errors for each posterior distribution, including every estimate of true flow for students and non-students were less than 5% of the standard deviation.

Through the interpretation of posterior characteristics for the data assessment parameters, one can make statements of the estimated effect for each parameter on true flow. As such, the first stage of the detailing of results is to analyse how the true flow is affected by the posterior estimates of the parameters. Comparisons between the priors and posteriors are also made to determine where the model is sensitive to prior assumptions. The main conclusions from this guide the sensitivity analysis later in the chapter.

The posterior characteristics of the data assessment parameters are detailed in table 7.1. Note carefully that they are expressed on the linear scale for ease of interpretation. Overall, where the priors are expressed with a relatively high level of certainty, the data is drawn towards the priors. For example, the posterior median and interquartile range (in brackets) for $def^{IPS.S}$ is 0.99 (0.98, 1.00) and for $def^{IPS.W}$ is 1.01 (1.00, 1.02). With the median values close to 1 on the linear scale, for the definition assessment criteria, the posterior indicates that the IPS is a good reflection of true flow.

This is also the case for the IPS bias terms. There are, however, slightly larger interquartile ranges in the posterior estimates. This is a reflection of the higher level of uncertainty expressed in the prior precision terms for IPS bias which, for both students and non-students is set at 1613; whereas, the IPS definition was elicited with a precision hyperparameter of 3246.

Where priors are elicited and the judgment is that there is undercount or overcount in a given data in relation to true flow, this is generally reflected in the posterior estimates. However, the posterior estimates for these parameters are, in the main, the most uncertain.

The largest amount of uncertainty in the posterior estimates can be found in the EU IPS coverage parameters. The posterior median and interquartile ranges for $cov^{IPS.EU.S}$ is 1.51 (1.42, 1.60) and for $cov^{IPS.EU.W}$ is 1.84 (1.72, 1.96). The assessment of the IPS data led to a prior that inflates the data towards true flow. This is still evident in posterior estimates as they are greater than 1 on the linear scale, albeit with a large level of uncertainty.

Model	Parameter	2.5%	25%	Median	75%	97.5%
Student Data Assessment Model	$def^{IPS.S}$	0.96	0.98	0.99	1.00	1.03
	$cov^{IPS.S}$	0.96	0.99	1.00	1.01	1.03
	$cov^{IPS.EU.S}$	1.27	1.42	1.51	1.60	1.78
	$bias^{IPS.S}$	0.93	0.96	0.98	1.00	1.03
	def^{HESA}	1.07	1.10	1.12	1.14	1.18
	cov^{HESA}	1.39	1.43	1.45	1.47	1.51
Non-Student Data Assessment Model	$def^{IPS.W}$	0.97	1.00	1.01	1.02	1.04
	$cov^{IPS.W}$	0.97	1.00	1.02	1.04	1.07
	$cov^{IPS.EU.W}$	1.50	1.72	1.84	1.96	2.23
	$bias^{IPS.W}$	0.97	1.00	1.01	1.03	1.06
	def^{DWP}	1.01	1.03	1.04	1.06	1.10
	$def^{DWP.EU}$	1.15	1.21	1.25	1.29	1.39
	cov^{DWP}	0.63	0.68	0.71	0.75	0.82

Table 7.1 Posterior Estimates of Data Assessment Parameters Under Informative Priors

The non-EU IPS coverage posteriors for students and non-students have a much lower level of uncertainty, where the median and interquartile range for $cov^{IPS.S}$ is 1.00 (0.99, 1.01) and for $cov^{IPS.W}$ is 1.02 (1.00, 1.04). This is a reflection of the prior elicitation for each of the IPS coverage parameters as there is a much greater level of certainty in the non-EU coverage priors. The effect on the posterior estimates of increasing precision in the priors, for the IPS coverage parameters, for both EU and non-EU flows, is tested in the sensitivity analysis.

As documented throughout the thesis, one of the main flows of UK immigration in the 2000s comprises Polish citizens, following the expansion of freedom of labour movement in the EU. Consequently, the true flow value of non-student Polish citizens is considered first. Figure 7.3 plots the posterior estimates of true flow for Polish non-students. Note carefully, that for each citizenship flow analysed in section 7.4 the median and the interquartile range (50% credible interval) from the posterior is plotted in comparison to the data used in the model.

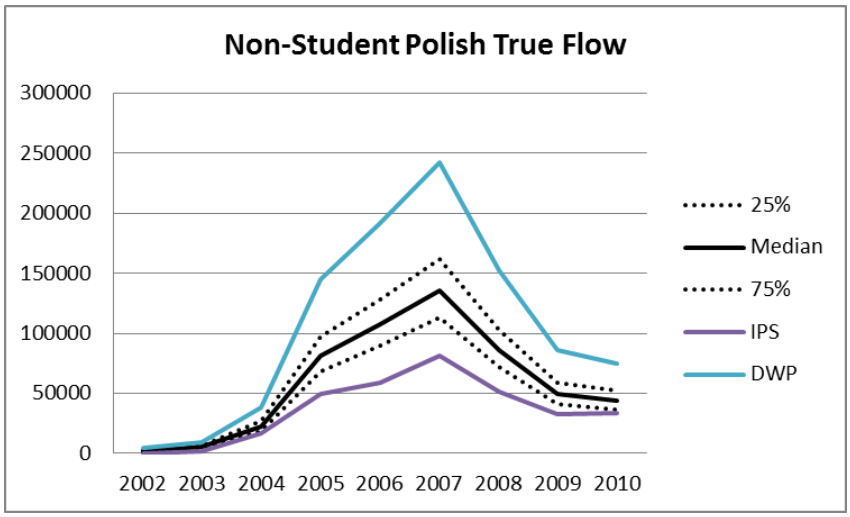

Figure 7.3: Comparison of Posterior Estimate of Non-Student Polish True Flow with the Data

The change in Polish immigration over time is as one would expect. Furthermore, through an estimate of a credible interval of true flow that is lower than the DWP data used, it seems that the overcount of short term migrants has been taken into account. Here, one of the main distortions of true flow is the overcount of short term migrants in the DWP data. This is taken into account by the prior specified for $def^{DWP.EU}$ which has a median and interquartile range of 1.19 (1.16, 1.22). The posterior seems to be sensitive to this prior judgement with a median and interquartile range of 1.25 (1.21, 1.29).

From the data audit in chapter 3 we know that there is an overcount of EU immigration in the DWP, due to short term migrants registering for a NINo, and the posterior estimate of this is 25%. In chapter 2, and then throughout the thesis the increased effect on immigration of the expansion of the freedom of labour movement is considered. A large number of these migrants will included in the DWP data, as such, the effect of changing the uncertainty in the $def^{DWP.EU}$ prior on true flow is determined in the

sensitivity analysis.The DWP data, on average, is approximately 1.75 times greater than the posterior median estimates of true flow for non-student EU countries. Figure 7.4, where the posterior estimate of immigration of Swedish non-students is displayed, is a further example of the DWP data overcounting EU non-student true flow. The credible interval of true flow is more stable than the IPS estimate and in general follows the pattern exhibited in the DWP data.

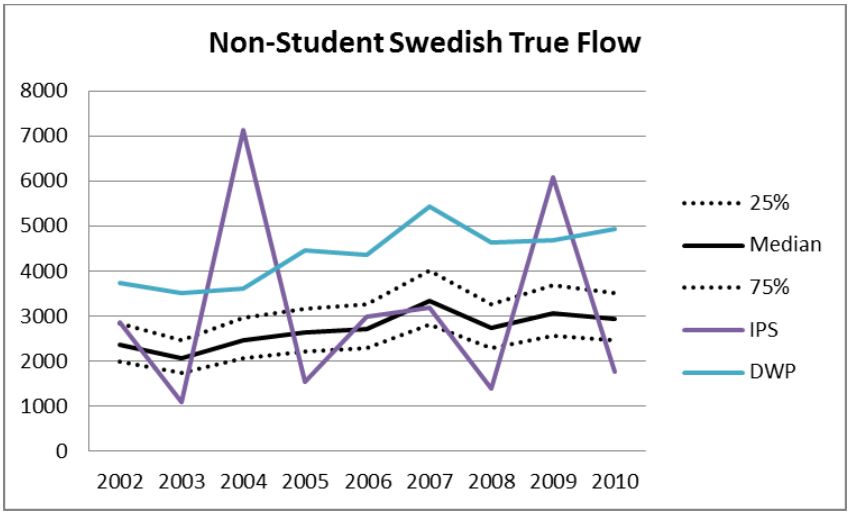

Figure 7.4: Comparison of Posterior Estimate of Non-Student Swedish True Flow with the Data

In comparison, for non-EU students the DWP data is only, on average, approximately 1.5 times greater than the posterior median of true flow. Figure 7.5, a plot of non-student true flow from the Philippines, illustrates this. The posterior credible interval is relatively closer to the DWP data than for EU flows.

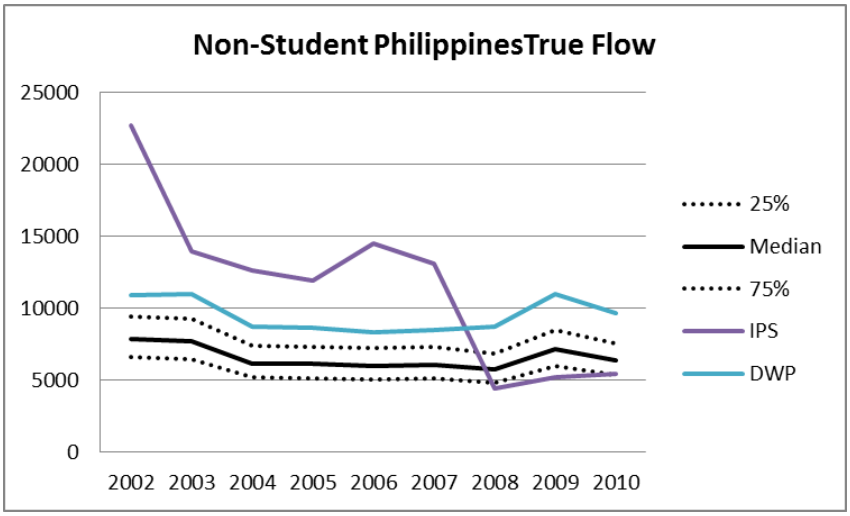

Figure 7.5: Comparison of Posterior Estimate of Non-Student Philippines True Flow with the Data

The median and interquartile range for the def^{DWP} prior is 1.05 (1.03, 1.06), which the posterior seems sensitive to with computed median values of 1.04 (1.03, 1.06). The judgement that there are fewer short term migrants distorting true flow for non-EU flows than EU flows is reflected in the posterior estimate being closer to the DWP data.

Finally, for the non-student part of the model, a plot of the posterior true flow estimate for Nigerian non-students is presented in figure 7.6. Here the posterior credible interval of true flow generally follows the pattern of the DWP data. With fewer short term migrants travelling to the UK from outside of the EU, due to the friction of distance and barriers to migration, the posterior median in comparison to an EU flow is relatively close to the DWP data. However, in 2006 the IPS estimate declines. The result of this is a relative divergence between the true flow estimate and the DWP data as the true flow posterior takes into account the data from both sources.

Where the DWP and non-student IPS data are a closer match the level of uncertainty in the posterior true flow is lower. This is evident in figure 7.6 and is also evident in figure 7.3 for Polish citizens. Generally, this is the case for each citizenship with the interquartile range of true flow positively correlated with the difference between the IPS and DWP data. A simple Pearson correlation coefficient for these two measures is significant and has a value of 0.883.

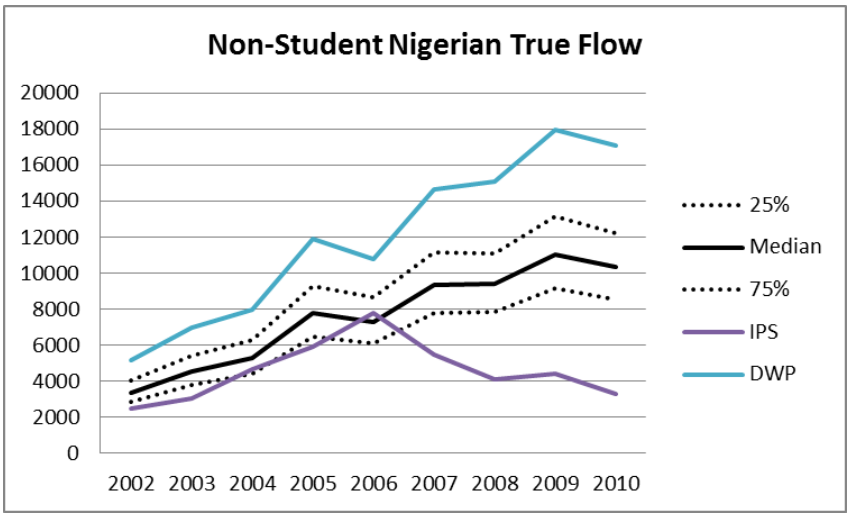

Figure 7.6: Comparison of Posterior Estimate of Non-Student Nigerian True Flow with the Data

Intuitively this makes sense. Where we have two sources of strong evidence, in this case data, that corroborate then the level of uncertainty should be lower than when this evidence is contradictory.

A further distortion of the DWP data that is taken into account in the non-student part of the model relates to the coverage data assessment criteria. This parameter assesses the effect on true flow of children not being included but students who work part time being counted in the DWP data. The distortion to true flow of cov^{DWP} needs to be taken into account to aid our understanding of the patterns exhibited in figures 7.3, 7.4, 7.5 and 7.6.

Uncertainty in the posterior estimate of the DWP coverage parameter is relatively high, however. The median and interquartile range for cov^{DWP} is 0.71 (0.63, 0.82). The posterior estimates that there is on average a 29% overcount in the DWP data in relation to true flow. However, for the prior it was judged that there is undercount caused by the exclusion of children from the DWP data and overcount as a result of students working part time being included, as this part of the model is for the estimate of non-students. The number of children excluded is judged to be higher than the number of students working part-time. With no concrete evidence of this, though, a relatively high level of uncertainty was elicited for the prior, with a median value and interquartile range of 1.10 (0.98, 1.23).

It would seem that the uncertainty in the prior has led to the posterior being relatively insensitive to this judgement. Ideally more certain prior information is required on the effect of this distortion of true flow as we know for certain, from the data assessment in chapter 3, that both the undercount of children and overcount of students is present in the DWP data. However, we do not know by how much and the certainty with which this judgement is applied is also up for debate. Furthermore, the posterior median of 0.71 for cov^{DWP} may be balancing the posterior median of 1.84 for $cov^{IPS.EU.W}$.

In the sensitivity analysis, the effect of tightening the DWP and then IPS EU prior certainty for coverage is detailed. For the DWP data, this has the effect of expressing with more confidence in the distortion caused by the undercount of children and for the IPS data this has the effect of expressing with more confidence the distortion caused to true flow by under-sampling at regional airports. From this sensitivity analysis it should be possible to determine any secondary effects tightening the prior certainty has.

For the student model the distortions of HESA, in relation to true flow, are taken into account by the definition and coverage parameters. The prior judgement is that there are students included in the HESA data that do not complete their first year of study and are likely to have not been usually resident in the UK for 12 months. The posterior median

and interquartile range for def^{HESA} is 1.12 (1.10, 1.14). Here the posterior is estimating that there is a 12% overcount in the HESA data for the definition assessment criteria in relation to true flow. There is also quite a high level of certainty in the posterior estimate of, suggesting that true flow

For EU and non-EU students, the posterior distribution of true flow estimates a higher level of immigration than both the HESA and IPS data. The posterior credible interval for the student true flow of French citizens is plotted in figure 7.7 for illustrative purposes. As with the non-student flows the posterior estimate is plotted on the same chart as the data used in the model.

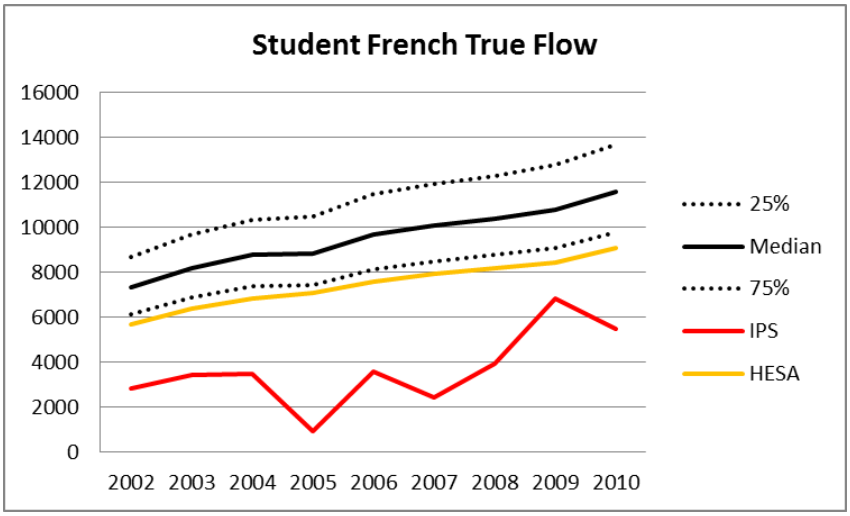

Figure 7.7: Comparison of Posterior Estimate of Student French True Flow with the Data

In general, across different citizenships, the true flow credible interval estimates a higher level of immigration than the HESA and IPS data. It also tends to follow the same pattern as the HESA data. The posterior median and interquartile range for cov^{HESA} is 1.45 (1.43, 1.47), which has quite a level of certainty. This could help explain the relatively high level of true flow in relation to the HESA data which is judged to have significant undercount.

Similar characteristics are also evident in the posterior estimate of true flow for Chinese citizens displayed by figure 7.8. The posterior credible interval is estimated to be greater than the reported HESA data and as with non-students, where the data are similar the interquartile range of true flow is tighter.

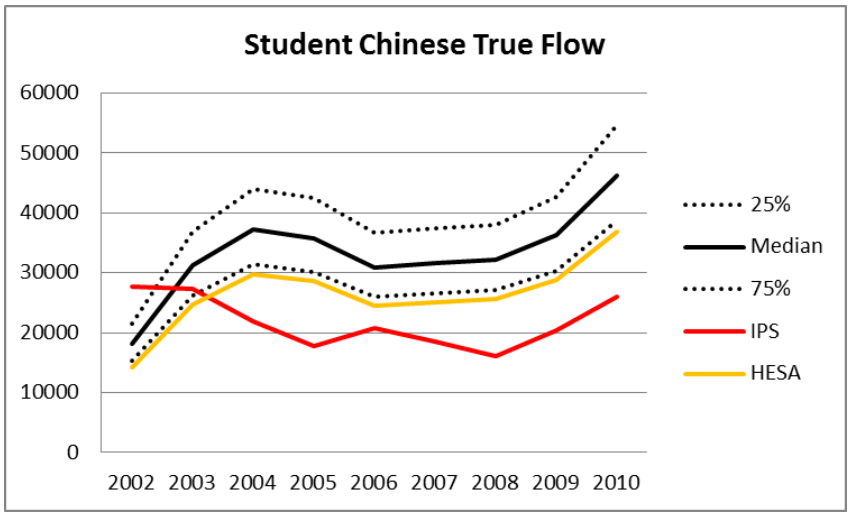

Figure 7.8: Comparison of Posterior Estimate of Student Chinese True Flow with the Data

Finally the posterior credible interval of Irish true flow is displayed by figure 7.9. Here the IPS estimate is that there was very few student migrants between 2002 and 2007. However, this does not seem to affect the estimate of true flow, with the posterior still seemingly strongly influenced by the HESA data and the cov^{HESA} prior.

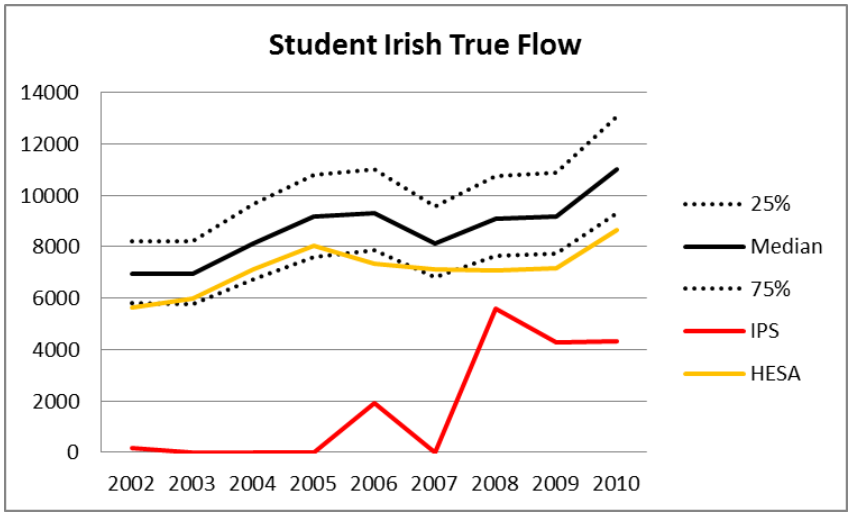

Figure 7.9: Comparison of Posterior Estimate of Student Irish True Flow with the Data

In the model specification the student and non-student true flow terms are summed together to provide one estimate of immigration with a coherent expression of uncertainty. Figure 7.10 plots the posterior credible interval of total Canadian immigration. This can only be compared to the total from the IPS, as administrative data – DWP and HESA are taken from different sources.

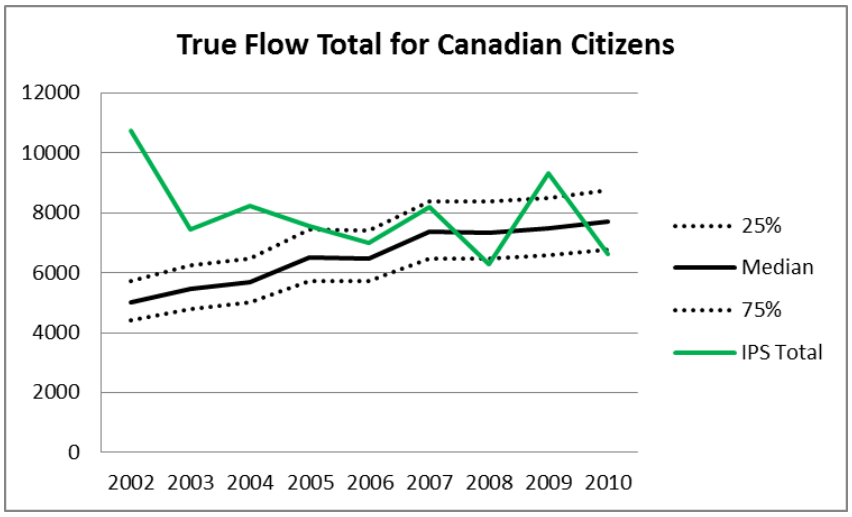

Figure 7.10 Comparison of Posterior Estimate of Total Candian True Flow with IPS Data

The stability of the estimate of Canadian immigration is a common feature in all the total true flow plots. Furthermore, the credible intervals seems a realistic representation of the uncertainty inherent in the publicly available immigration statistics and our prior knowledge elicited through the priors.

Finally a brief comparison between the posterior estimates of the Bayesian Log-linear model and Data Assessment Model is made. Firstly, the true flow of Polish citizens is considered. For the Bayesian Log-linear model the computation using $\tau_{\lambda_t^S}^{[3]}$ as the precision term is compared to the Data Assessment Model Results in figure 7.11 below.

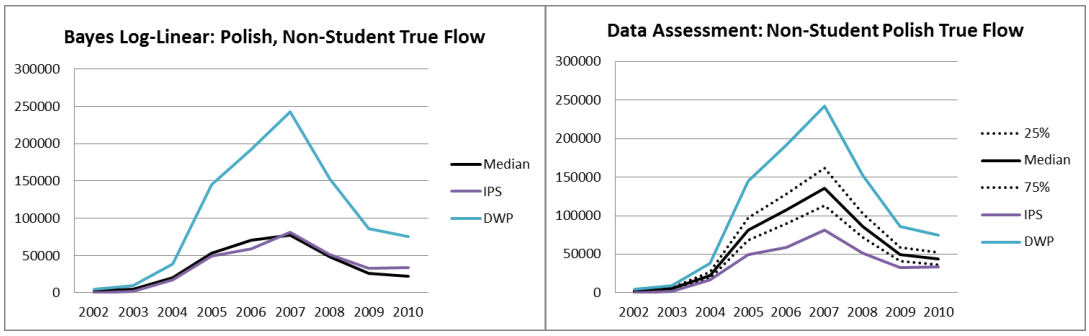

Figure 7.11 Comparison of the Bayesian Log-linear Model and Data Assessment Models True Flow Posteriors for Polish Non-Students

Both charts are displayed using the same scale to allow for comparison. Also note that only the posterior median has been plotted for the Bayesian Log-linear estimate of Polish true flow, as the interquartile range is so narrow it renders the graph unreadable. Unlike the Bayesian Log-linear model, the estimate of true flow taken from the Data Assessment Model on the right is not constrained to fit the IPS margins. This allows

posterior estimates of true flow which, where appropriate, are allowed to exceed the IPS totals.

For students, posterior estimates of the true flow of Chinese citizenships is plotted in figure 7.12. Similar to the example detailed above of the non-student true flow posteriors of Polish migration, it is clear that the posterior estimate of true flow is unconstrained by the IPS margins. We know from the data audit in chapter 3 that there is under count in the HESA data as a result of a lack of coverage for FE students and languages colleges. This judgement cannot be taken into account within the marginal constraints of the Bayesian log-linear model; whereas, in the Data Assessment Model, this judgement has been modelled explicitly through the data assessment equations.

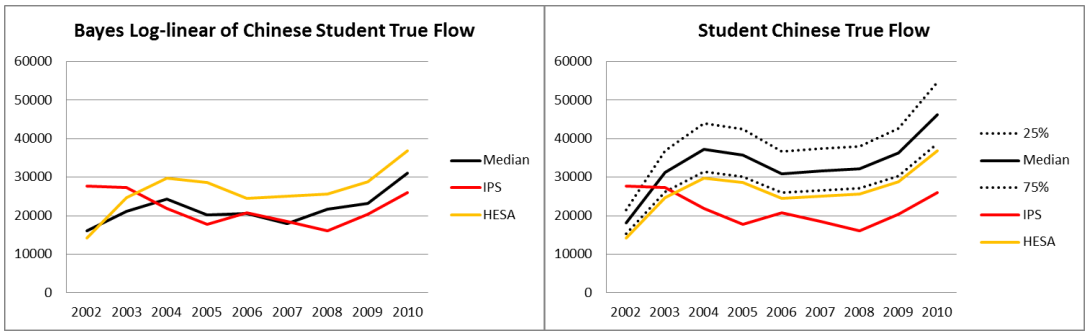

Figure 7.12 Comparison of the Bayesian Log-linear Model and Data Assessment Models True Flow Posteriors for Chinese Students

7.4.2 Sensitivity Analysis – IPS Coverage

As outlined in chapter 6.4.2 the elicitation of the priors for the Data Assessment Model is of significant consequence. The model, unlike the Bayesian Log-linear model, is not marginally constrained to fit given values. As a result it is important to test the sensitivity of the posterior estimates to changing the prior assumptions. In this section various judgements about the distortion of true flow are changed from the original prior elicitation carried out in chapter 6.4.3. This is guided from the results of the informative prior model computation.

Following the analysis of the posterior estimates under informative priors, it is clear that a large amount of uncertainty from the IPS coverage priors is propagated into the posterior estimate of true flow. In this sensitivity analysis each of the IPS coverage priors

are re-elicited with a higher level of uncertainty, by bringing either the upper or lower tail of the distribution closer to the median value.

The median values of the priors remain the same, as this is the best judgement of the quantity of the distortion of true flow. However, in this sensitivity analysis we are testing the effect on the posterior estimates if we express this judgement with a higher level of certainty. In the case of the IPS coverage parameters, this means have a greater amount of faith in the complete survey coverage of non-EU citizens. However, with regard to EU citizens, for both students and non-student, increasing the certainty in the prior means having a greater level of certainty that there is undercount caused by lack of survey coverage at regional and smaller airports.

For each of the priors, the new level of uncertainty was elicited using the quantile method outline in chapter 6.4.1. For $cov^{IPS.S}$ and $cov^{IPS.W}$ the medians are kept at 1 on the linear scale and the upper quantiles are elicited as 1.035 and 1.025 respectively, providing prior distributions of $cov^{IPS.S} \sim N(0, 3246)$ and $cov^{IPS.W} \sim N(0, 6300)$ on the log scale. For $cov^{IPS.EU.S}$ and $cov^{IPS.EU.W}$ the prior medians are kept at 1.2 and 1.4 on the linear scale respectively. The quantiles are then elicited by reducing the interval between the median and the lower quantiles by half. The lower quantiles are elicited as 1.1 and 1.2 on the linear scale respectively, providing prior distributions of $cov^{IPS.EU.S} \sim N(0.18, 507)$ and $cov^{IPS.EU.W} \sim N(0.33, 161)$.

All other priors in the model remain informative, as elicited in chapter 6. The model is computed with a 7000 (thinned every 100th) burn in, after which convergence is achieved and then a further 8000 iterations (also thinned for every 100th), providing accurate posterior calculations.

Having obtained the posterior estimates, it is apparent that any secondary effects in the model are very small. As such, any sensitivity to posterior true flow estimates are mainly a result of the changes in the IPS coverage priors. The posterior estimate of $cov^{IPS.EU.W}$ is of particular interest, because of the judged undercount of EU migrants at regional airports.

Figure 7.13 shows the posterior plots of $cov^{IPS.EU.W}$ estimated under informative (left-hand plot) and then sensitivity priors (right hand plot). Please note these plots are taken directly from the computation, and as such the x-axis is on the log-scale. With the tighter coverage priors the median and interquartile range are 1.58 (1.51, 1.65) on the linear

scale. This is closer to the elicited prior median of 1.4, than when the model was estimated with informative prior. This is expected given the increased level of certainty expressed in the prior during this sensitivity analysis. Increasing the prior certainty for $cov^{IPS.EU.W}$ has also had the effect of reducing the uncertainty in the posterior estimate, which can be seen below.

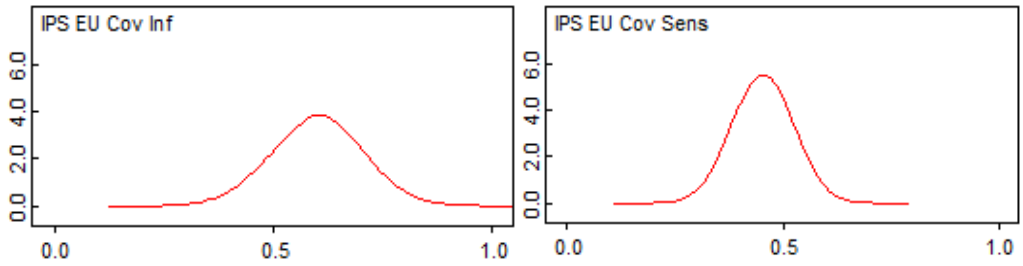

Figure 7.13: Comparison of Posterior for $cov^{IPS.EU.W}$ Under Informative Priors and then Sensitivity

This increased level of certainty in the $cov^{IPS.EU.W}$ posterior, only has a small effect on the estimate of true flow. Figure 7.14 below uses Sweden as an example. The true flow posterior credible interval is around 10% lower under sensitivity than under informative priors.

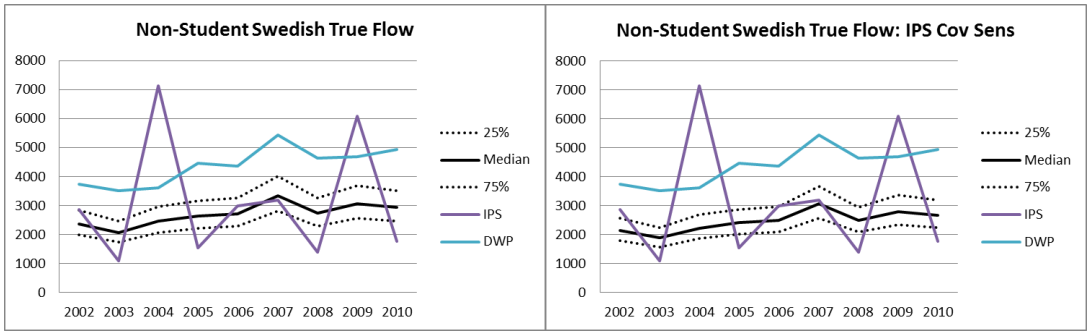

Figure 7.14 Sensitivity of the Data Assessment Model to More Certain IPS Coverage Priors, for Posterior True Flows of Swedish Non-Students

For the $cov^{IPS.EU.S}$ the posterior is sensitive to the prior assumptions of the sensitivity with a smaller median and tighter interquartile range of 1.27 (1.23, 1.30) in comparison to the model estimate under informative priors (see table 7.1). However, the true flow posteriors are largely insensitive to this, as illustrated by figure 7.15 below which compares the posterior estimates for French students under informative on the left and under the sensitivity on the right.

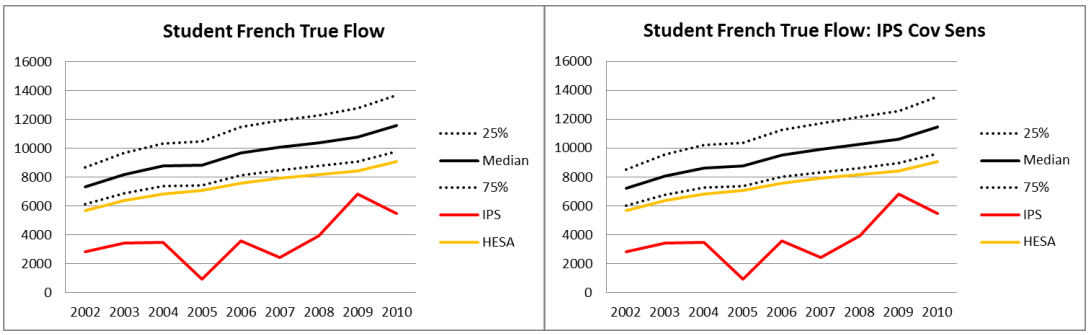

Figure 7.15: Sensitivity of the Data Assessment Model to More Certain IPS Coverage Priors, for Posterior True Flows of French Students

7.4.3 Sensitivity Analysis – HESA and DWP Coverage

The model is re-estimated with more certain HESA and DWP coverage priors to test the sensitivity of the posterior estimates to there being a greater level of certainty in the judgements made about the HESA and DWP is distorted in relation to true flow. The new priors were elicited using the quantile method. For cov^{HESA} and cov^{DWP} the medians remain 1.47 and 1.1 on the linear scale. For the HESA data the upper quantile (97.5) is used and is elicited as 1.50 and for the DWP data the lower quantile (2.5) is elicited as 1, both on the linear scale. The quantiles are elicited by reducing the interval between the median and either the upper or lower quantile by half. This results in priors of $cov^{HESA} \sim N(0.38, 8879)$ and $cov^{DWP} \sim N(0.095, 423)$. All other priors in the model remain informative and the computation, with regard to the burn in and number of iterations, is the same here as for section 7.4.2.2.

For cov^{DWP} the posterior is sensitive to the tightening of the prior. The median and interquartile range of the computed posterior is 0.94 (0.92, 0.97). The posteriors under informative priors (left-hand plot) and under sensitivity (right-hand plot) are displayed in figure 7.16. The sensitivity analysis has led to a posterior estimate with more certainty and a closer match to the elicited prior median.

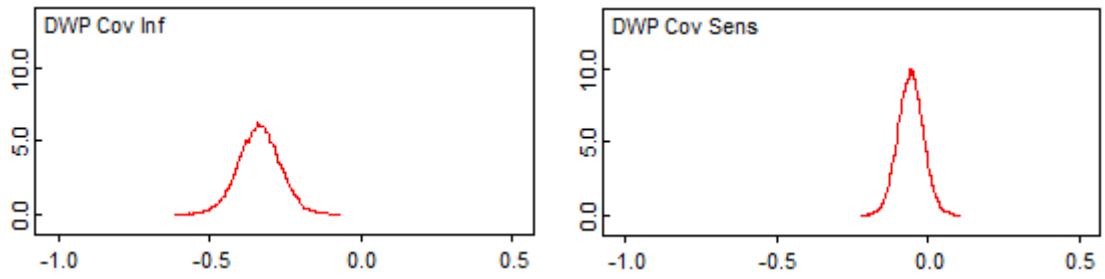

Figure 7.16: Comparison of Posterior for cov^{DWP} Under Informative Priors and then Sensitivity

There is also a secondary effect, where changing the prior for a given parameter also results in posterior sensitivity for a different parameter, which is evident for the median and interquartile range for $cov^{IPS.EU.W}$, estimated as 2.15 (2.02, 2.29). This is a significant change in comparison to the posterior computed under informative priors; however, the large interquartile range might limit the effect of this data assessment parameter on the true flow posteriors. Also the informative prior for $cov^{IPS.EU.W}$ is quite uncertain, so one would expect secondary effects here as the prior judgement is based purely on subjective judgement and on any concrete evidence.

The true flow posterior credible interval for Nigerian non-students, under informative priors (left) and sensitivity (right), is plotted side by side in figure 7.17. It is clear that the sensitivity of the DWP coverage posterior to a tighter prior leads to an increased estimate of true flow. This increase is approximately 20% and suggests that more information is needed on the coverage of the DWP data.

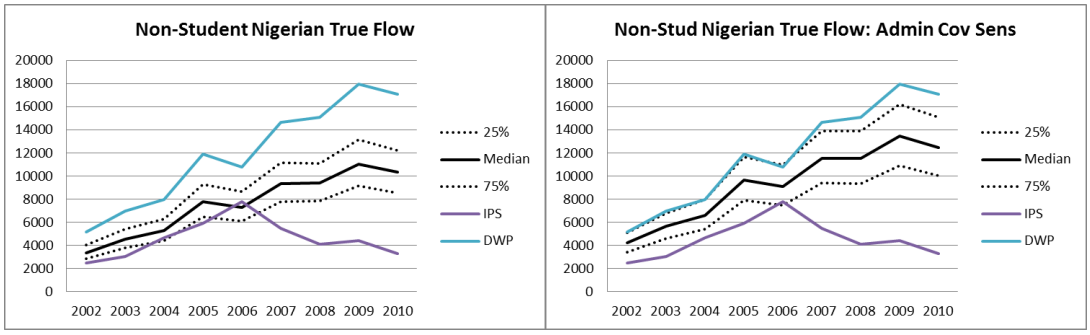

Figure 7.17: Sensitivity of the Data Assessment Model to More Certain DWP and HESA Coverage Priors, for Posterior True Flows of Nigerian Non-Students

As the secondary effect is for an EU-specific data assessment parameter, it is necessary to look at an example of an EU non-student flow. In figure 7.18 below there is a comparison of the true flow posteriors for non-student Italian migration under informative priors (left) and sensitivity (right). The sensitivity in the true flow posterior is also clear. There is an increase in true flow of roughly 30% as a result of the sensitivity analysis. This may be from the secondary effect detailed earlier, however.

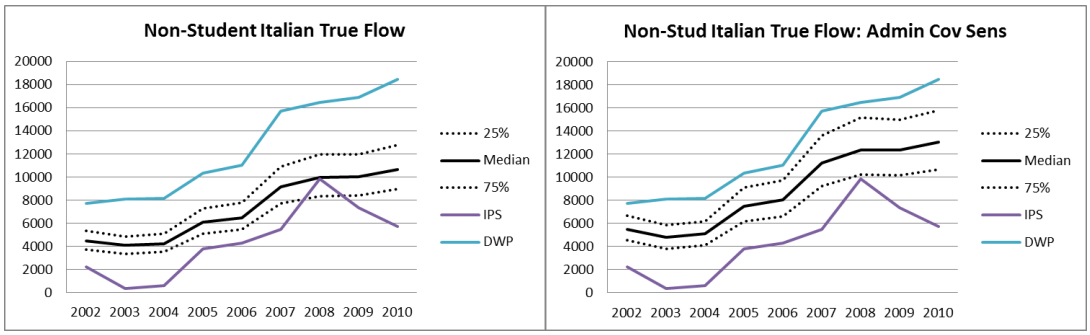

Figure 7.18: Sensitivity of the Data Assessment Model to More Certain DWP and HESA Coverage Priors, for Posterior True Flows of Italian Non-Students

The cov^{HESA} posterior is largely insensitive to the tightening of the priors. This is expected as under informative priors there was a close prior/posterior match. The median and interquartile range of the posterior under sensitivity is 1.46 (1.45, 1.47). The posterior estimates of the other student parameters are largely insensitive, and there are no significant secondary effects.

However, when one looks at the posterior estimates for true flow, there is sensitivity as a result of tightening the HESA and DWP coverage parameters. An example is given below in a comparison of Chinese non-student true flows under informative priors and sensitivity in figure 7.19.

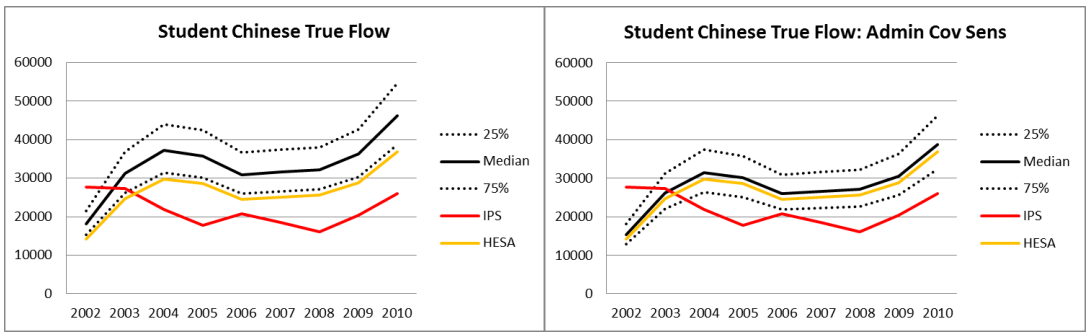

Figure 7.19: Sensitivity of the Data Assessment Model to More Certain DWP and HESA Coverage Priors, for Posterior True Flows of Chinese Students

Referring back to figure 5.2 in chapter 5, the diagram shows that all of the parameters in the model are linked. This means that a change in the prior assumption for non-students could have a secondary effect in the student model. This is a limitation to estimating students and non-students in one model.

As a result of the sensitivity analysis for the HESA and DWP coverage parameters, it is clear that more information is needed about the undercount of children and overcount of students in the DWP data. There is a significant effect on the posterior estimates of true flow for both students and non-students.

7.4.4 Sensitivity Analysis – HESA and DWP Definitions

The final sensitivity analysis is the elicitation of more certain HESA and DWP definition priors. This is to check posterior sensitivity to increasing the certainty of the judgement of overcount in the HESA and DWP definition priors. Here the certainty is increased to greater extent than in the previous sensitivity analyses that focussed on the coverage parameters. The main reason for this, is that we have concrete evidence, taken from student drop-out rates and the prior elicited by Raymer et (2013) from a panel experts and the aim is to test the sensitivity of the posterior if we strongly believe in this prior information as an indication of the distortion of true flow.

The prior distributions for these parameters are assumed to be gamma distributed. The new level of certainty of each prior is elicited using the quantile method which, as outlined in chapter 6.4.1, is based on assumptions of normality. For def^{HESA} the prior median remains set at 1.01 on the linear scale and for $def^{DWP.EU}$ and $def^{DWP.EU}$ the prior medians are kept at 1.19 and 1.1. Here, as with each of the stages of sensitivity analysis, the quantiles are elicited by reducing the interval between the median and the upper (97.5) or lower (2.5) quantile. The HESA upper quantile is elicited as 1.11 results in a prior distribution $def^{HESA} \sim \Gamma(511, 4895)$. The EU DWP prior is elicited using the upper quantile, which is set at 1.21 and the non-EU lower quantile is set at 1.045. This results in prior distributions of $def^{DWP.EU} \sim \Gamma(502,\ 2636)$ and $def^{DWP} \sim \Gamma(401, 8226)$.

Computation is carried out in the same way as the previous sensitivity analysis, with checks for convergence on each parameter carried out and Monte Carlo Errors checked for each parameter.

The posterior estimates are largely insensitive to changes in the definition priors. The only sensitivity that is evident is in the $def^{DWP.EU}$ parameters where the posterior has a median and interquartile range of 1.21 (1.20, 1.22) in comparison to 1.24 (1.21, 1.29).

However, when the true flow posteriors are compared to the posteriors computed under informative priors there is very little difference for both students and non-students.

There are very small secondary effects in each of the other posteriors, and as such, it could be the case that the model simply balances the sensitivity of the $def^{DWP.EU}$ and any effect on true flow with these smaller secondary effects. Thus cancelling out any sensitivity to the DWP EU definition prior.

The data audit and the results of the Data Assessment Model could provide a starting point for the ONS. When more evidence on the distortion of true flow, caused by various characteristics of data collection comes to light, the priors can be updated and our uncertainty of characteristics of immigration data and true flow adjusted accordingly.

Where there is strong evidence available on the distortion of true flow, the posterior estimates have been sensitive to this. For example, the posterior of the HESA coverage parameter indicates, with a reasonable level of certainty, that the estimate of true flow is sensitive to the undercount judgement of the prior. This prior is taken from visa applications for non-EU citizens. Another recommendation for the ONS is to carry out an audit of all international student data that is held by public institutions. This could then either be used to supplement the HESA data, to create a quasi-register of international students or to possibly update the priors in the model. The results of the true flow posterior estimates for students show that the HESA data is influential in providing a stable pattern of citizenship specific immigration over time in comparison to the relatively noisy IPS data.

7.4.4 Discussion and Limitations

Where data assessment parameters attempt to take into account more than one distortion of true flow, the posterior may not have the desired effect on true flow. This is the case for the cov^{DWP} where the prior judgement is based on an undercount of children and an overcount of students who work part time. If one is to carry out a data assessment using a general framework, rather than simply specifying very specific priors, then this type of problem is encountered. In terms of the overall approach of the research, however, including both the under and overcount in the same parameter (and then testing for

sensitivity) is an adequate compromise between the conceptual clarity of the data assessment criteria and the practical challenges of estimating the model.

Priors with a reasonably high level of certainty are needed for purposes of identifiability in the model. Also, the model does not converge if vague priors are specified. Furthermore the level of certainty with which a prior is elicited has a strong effect on the posterior estimate in this model. What this means is that secondary effects tend to be found in data assessment parameters that have a higher level of uncertainty. This could have the effect of cancelling out a consideration of the distortion true flow included in a different prior. Within this framework there is not a straight forward solution to this problem, as uncertain priors can be justified when our assessment of the data calls for some uncertainty to be introduced to model, for example, in the assessment of the EU specific survey coverage of the IPS.

7.5 Conclusions

From the results in this chapter, it has become clear that in order to include the assessment of the available data explicitly in a statistical model, then there is a need to move beyond the constraints of the log-linear framework.

The Bayesian log-linear framework effectively utilises the marginal strengths of the IPS in terms of its close match to true flow with regard to definition and combines this with the accuracy of the accurate patterns of the HESA and DWP data. Furthermore, when confronted with auxiliary data, used in priors for the main effects and specified with a high level of certainty, the data remains dominant. This suggests that we are not learning anything more from including the Flag 4 data and Census data as alternative marginal effects.

Uncertainty in the posterior distributions estimated for the Bayesian Log-linear model, however, remain strongly influenced by the stochastic error in the model, rather than an expression of uncertainty about the estimate of true flow. If the data is made available, future research should include the survey weights, rather than the IPS population- level estimates. The uncertainty in the posterior would then include a consideration of the random variation of the sample survey and the judgements made for the main effects priors in the model.

Removing the constraints of the log-linear framework is a more appropriate approach when there is a need to model explicitly subjective judgements of what we know about data collection. However, the model estimates are more sensitive to the priors, which makes elicitation very important. Furthermore, the presence of secondary effects, also means care needs to be taken when drawing conclusions about the sensitivity of the posterior true flow estimates to the subjective judgements of the data.

However, where there are priors that are based on concrete evidence, the posterior of the data assessment parameters have a relatively high level of certainty which seems to be reflected in comparisons of true flow to the sources of data. Stable patterns over time are produced, and where we are more certain about under-count or over-count this is reflected in the posterior distributions of the respective true flows.

With regard to the assessment parameters, it is clear, from the uncertainty present in the posteriors and from the sensitivity analysis that there needs to be more research into the survey coverage of IPS, especially for EU citizenships. Furthermore, for the DWP data more information is required on the number of students who work part time and the number of children who are not included. Any new information can then form the basis for further data assessment, or can be used as priors in a similar model.

Chapter 8 Conclusion

8.1 Summary of Main Contributions

There are two main aims in this thesis. The first is to use statistical models to make better use of all publicly available data and information in the estimation of UK immigration. The second is to understand better the amount and specific sources of uncertainty in the publicly available data. In order to fulfil these aims three statistical models have been developed, applied and the results have been interpreted. Each model is based on the data audit and assessment carried out in chapter 3, to ensure appropriate model specification and interpretation of results.

The data assessment criteria proposed in chapter 3 are a key contribution to improving our understanding of the nature of uncertainty in UK immigration statistics. The proposed framework provides the basis for research into the quantity, nature and the uncertainty of the distortion to the measurement of true immigration flows caused by the process of data collection. The table summary of the data assessment criteria for each publicly available source of data is then crucial to the development and application of the three modelling approaches outlined in chapters 4 and 5.

The log-linear model outlined in chapter 4 reflects the current state of the art with regard to UK immigration estimation. The model succeeds in combining the IPS totals with the more reliable patterns exhibited in the HESA and DWP administrative data. However, a limitation of this model is that judgements of uncertainty cannot be modelled within a standard log-linear framework, and the estimates are constrained to the IPS marginal totals.

The log-linear framework in chapter 4 is extended in chapter 5 to include judgements of auxiliary marginal data. One of the contributions of chapter 6 is proposing a method for eliciting prior distributions for the main effects parameters in the Bayesian log-linear framework. There are few examples of prior elicitation of this kind, in an applied setting in the literature. As such, the elicitation of the prior medians for the Bayesian Log-linear main effects parameters, is a small contribution to applications where subjective judgements are elicited for Bayesian log-linear models in general.

A limitation of the Bayesian Log-linear approach, however, is that the estimation of uncertainty is again simply a reflection of the stochastic error in the model. Future research should apply this method using the survey weights, which have not been available for this study. The posterior estimates would then include an additional expression of the uncertainty from the IPS sampling error. Furthermore, the Bayesian Log-linear model is also constrained to the log-linear margins of the IPS. This does not allow for the coverage and bias data assessment criteria to be considered in the model specification. In order to explicitly consider all of the data assessment criteria, an unconstrained model is required.

The Data Assessment Model, outlined in the second part of chapter 5, helps address the second aim of the thesis. It is unconstrained, which addresses the key limitation of the Bayesian Log-linear model. Importantly, however, it is developed with the data assessment criteria, outlined in chapter 3, explicitly in mind. Through taking a fully Bayesian approach, the results include coherent expressions of uncertainty through the estimate of posterior distributions based on the information in the data and our subjective judgement of the data distorts true flow.

As such, the importance of prior elicitation for the Data Assessment Model cannot be underestimated. The prior elicitation for the Data Assessment Model in chapter 6, is the first of its kind for UK immigration. The developed quantile-based elicitation provides elicited prior judgements that relate directly to the data assessment criteria, which can then be fitted as a probability distribution and used as prior information.

Moreover, from the results and sensitivity analysis of the Data Assessment Model in chapter 7, it is clear that where prior judgements have a low level of uncertainty, they are affected by sensitivity in the posterior estimates, which has an impact on the estimate of true flow. Of course this means that care needs to be taken in the elicitation of prior judgements. Future research is required to strengthen the process of prior elicitation further. An approach similar to that of Bijak and Wiśniowski (2010), which elicits prior information from a panel of experts, using a Delphi survey could be used. Using the quantile approach, outlined in chapter 6 however, may not be appropriate when consulting the expert judgement of non-statisticians. As an alternative, the approach outlined by Wiśniowski, Bijak and Shang (2014), where experts were asked to state their forecasts of migration and to place a 'bet' to indicate the probability and certainty of their forecast, could be applied to judgements of UK immigration data.

Finally, the conclusions of chapter 7, with regard to the assessment of data, could be used in further research to update the data assessment carried out in chapter 3. The process of continually updating our uncertainty and prior judgement of UK immigration data and estimates, as a result of this and future research, would be a strong contribution to the practice of UK migration estimation.

8.2 Recommendations to the ONS for Data Collection

As outlined in chapter 4, the UK Statistics Authority, as a result of evidence submitted by Bjiak et al (2013) to the Public Administration Select Committee report on migration (PASC 2013), have suggested to Government that alternative sources of data should be used in the estimation of UK immigration (UKSA 2013). It is a recommendation of this thesis that the ONS, in the short term, look into the approach outlined in chapter 4 to fulfil the UKSA suggestions.

In the longer term, a further recommendation for the ONS is that they could explore the possibility of further research into some of the main findings of the Data Assessment Model. The main sources of uncertainty that were found in the results of this model come from the survey coverage of the IPS, for EU immigration flows. Furthermore, the posterior estimates are also sensitive to changes in the IPS coverage priors.

The prior judgement of the data is that there is an undercount of EU migrants as a result of lack of IPS survey coverage at regional ports and airports. However, there is not any useable evidence, about the data collection process, that one can use to express both the size of this distortion and our certainty about this judgement. The ONS have acknowledged that undercoverage of this kind may be an issue after having to adjust their mid-year population estimates. They found that there was a discrepancy of 250,000 in A8 migration between the IPS estimate of immigration and the 2011 Census estimate (ONS 2012 c).

It is recommended that there are two main ways the ONS can deal with the problem of underestimation of A8 immigration by the IPS. Firstly, they could increase the number of survey shifts at regional airports, thus increasing the survey coverage. Since 2008 there have been more such shifts (ONS 2012 d). However, the ONS state that to decrease their estimated confidence intervals of migration flows by half would take a four-

fold increase in the size of the sample; and thus a four-fold increase in the cost of the survey from the current £5 million per year.

As an alternative, the ONS could continue their work to include administrative data in estimates of immigration. Their Migration Statistics Quarterly Reports have started to detail the number of NINo registrations, using the same DWP data that has been applied in this thesis. Taking into account the data assessment criteria, the ONS should invest more time and resources into obtaining migration estimates from administrative data, particularly the DWP data. Research into gaining a better understanding of the effect of short term and circular migration on the DWP estimate of true flow would be invaluable.

More generally, a further recommendation is that the official statistics on immigration into the UK explicitly include assessments, quantifications and judgements of uncertainty in their production and dissemination of immigration statistics. Understanding and estimating the main sources of uncertainty in the data will not only help produce better estimates, but also will guide future work on improvements to data collection.

Bibliography

Abel, G. (2010) *International Migration Flow Table Estimation* PhD, University of Southampton

Abel, G. (2013) "Estimating global migration flow tables using place of birth data" *Demographic Research* 28 (18) pp. 505-546

Agresti, A. (2002) *Categorical Data Analysis* New Jersey: Wiley

Agresti, A. (2013) *Categorical Data Analysis* 3rd Edition New Jersey: Wiley

Agresti, A. & Finlay, B. (2009) *Statistical Methods for the Social Sciences* Cambridge:Pearson

Arango, J. (2000) "Explaining migration: a critical view" *International Social Science Journal* 52 pp. 283 – 296

Ballard, R. (2002) "The South Asian Presence in Britain and its Transnational Connections" in Ballard, R. Singh, H. and Vertovec, S. (eds) *Culture and Economy in the Indian Diaspora*, London: Routledge

Beach, L. and Scopp, T, (1967) "Intuitive Statistical Inferences about Variances" *Organisational Behaviour and Human Performance* 3 pp. 109-123

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J. and Hugo, G. (2002) "Cross National Comparison of Internal Migration: Issues and Measures" *Journal of the Royal Statistical Society A* 165 pp. 435 – 464

Bennett, R. and Haining, R. (1985) "Spatial structure and spatial interaction: modelling approaches to the statistical analysis of geographical data" *Journal of the Royal Statistical Society A* 148 pp. 1 - 36

Bernardo, J. (2003) "Bayesian Statistics" in Viertl, R. (ed) *Probability and Statistics, Encyclopaedia of Life Support Systems* Oxford: UNESCO

Bijak, J. (2010) *Forecasting International Migration In Europe, A Bayesian View* Dordrecht: Springer

Bijak, J. (2010 a) *Independent Review of Methods for Distributing International Immigration Estimates to Regions* Southampton: University of Southampton [Online] Available at: http://www.cpc.ac.uk/resources/downloads/review_international_migration.pdf [Accessed 20 September 2014]

Bijak, J. Disney, G. Lubman, S. Wisniowski, A. (2013) *Towards Reliable Migration Statistics for the United Kingdom*. Report in response to the House of Commons Public Administration Select Committee Call for Evidence on Migration Statistics. CPC University of Southampton

Bijak, J. & Wisniowski, A. (2010) "Bayesian Forecasting of Immigration to Selected European Countries by Using Expert Knowledge" *Journal of the Royal Statistical Society, Series A* 173 pp. 775 - 796

Blinder, S. (2014 a) *Briefing: Migration to the UK, Non-European Student Migration to the UK* Oxford: Migration Observatory, University of Oxford [Online] Available at: http://www.migrationobservatory.ox.ac.uk/briefings/non-european-student-migration-uk [Accessed 20 September 2014]

Blinder, S. (2014 b) *Briefing: Migration to the UK, Asylum* Oxford: Migration Observatory, University of Oxford [Online] Available at: http://migrationobservatory.ox.ac.uk/briefings/migration-uk-asylum [Accessed 20 September 2014]

Bloemraad, I., Korteweg, A. and Yurdakul, G. "Citizenship and immigration: multiculturalism, assimilation, and challenges to the nation-state" *Annual Review of Sociology* 34 pp. 153-179

Boden, P. and Rees, P. (2010) Using administrative data to improve the estimation of immigration to local areas in England *Journal of the Royal Statistical Society* 173: 707-731.

Borjas, G. (1989) "Economic theory and international migration" *International Migration Review* 23 pp. 457 – 485

Boyle, P. and Dorling, D. (2004) Editorial: the 2001 UK census: remarkable resource or bygone legacy of the 'pencil and paper era'? *Area* 36: 101-110

Boyle, P., Exeter, D., and Flowerdew, R. (2004) "The role of population change in widening the mortality gap in Scotland" *Area* 36 pp. 164 – 173

Brierley, M. , Forster, J., McDonald, J. & Smith, P. (2008) "Bayesian estimation of migration flows" in Raymer, J. & Willekens, F. (eds) *International Migration in Europe: Data Models and Estimates* Chichester: Wiley

Brooks, S. and Gelman, A. (1998) "General methods for monitoring convergence of iterative simulations" *Journal of Computation and Graphical Statistics* 7 pp. 434-455

Byron, M. (1994) *Post-war Caribbean Migration to Britain: the Unfinished Cycle,* Aldershot, U. K

Castles, S. and Miller, M. (2009) *The Age of Migration: International Population Movements in the Modern World* Basingstoke: Macmillan

Chatfield, C. (2002) "Confessions of a Pragmatic Statistician" *The Statistician* 51 pp. 1-20

Coleman, D. and Rowthorn, R. (2004) "The economic effects of immigration into the United Kingdom" *Population and Development Review* 30 pp. 570-624

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

Congdon, P. (2006) *Bayesian Statistical Modelling* Chichester: Wiley

Davey-Smith, G., Dorling, D., Mitchell, R. and Shaw, M. (2002) Health inequalities in Britain: continuing increases up to the end of the 20th Century *Journal of Epidemiology and Community Health* 56: 434-435.

De Beer, J., Raymer, J., van Wissen, L. and van der Erf, R. (2010) "Overcoming the problems of inconsistent international migration data: a new method applied to flows in Europe" *European Journal of Population* 26 pp. 459 – 481

DWP (2010) *National Insurance Number allocations to adult overseas nationals entering the UK* London: Department for Work and Pensions

European Union (2007) Article 3 of the European Parliament Regulation (EC) No. 862/2007

Freedman, D. (2005) *Statistical Models Theory and Practice* Cambridge: Cambridge University Press

Freeman, R. (2006) "People flows in globalisation" *Journal of Economic Perspectives* 20 pp. 145-170

Garthwaite, P., Kadane, J. & O'Hagan, A. (2005) "Statistical Methods of Eliciting Probability Distributions" *Journal of the American Statistical Association* 100 pp. 680 - 700

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004) *Bayesian Data Analysis* Bury St. Edmunds: St Edmundsbury Press

Gill, J. (2002) *Bayesian Methods: A Social and Behavioural Sciences Approach* London: Chapman and Hall

Goldstein, M. (2006) "Subjective Bayesian Analysis: Principles and Practice" *Bayesian Analysis* 3 pp. 403-420

Hamnett, C. (1994) "Social polarisation in global cities: theory and evidence" *Urban Studies* 31 pp. 401 - 424

Hansen, R. (1999) "The politics of citizenship in 1940s Britain: The British Nationality Act" *Twentieth Century British History* 10 pp. 67 - 95

Harvey, D. (1999) *The Limits to Capital* Oxford: Blackwell

Hatton, T. (2003) *Emigration from the UK, 1870 – 1913 and 1950 – 1998* IZA Discussion Paper No. 830

Hatton, T. (2005) "Explaining trends in UK immigration" *Journal of Population Economics* 18 pp. 719-740

Hatton, T. and Price, S. (1999) *Migration, Migrants and Policy in the United Kingdom* Centre for Economic Policy Research – Discussion Paper No. 1960

Hatton, T, and Tani, M. (2005) "Immigration and inter-regional mobility in the UK 1982 – 2000" *The Economic Journal* 115 pp. 342 - 358

Hatton, T. and Williamson, J. (1998) *The Age of Mass Migration: Causes and Economic Impact* Oxford: Oxford University Press

Hatton, T. and Williamson, J. (2002) "What fundamentals drive world migration" *Centre for Economic Policy Research - Discussion Paper No. 458*

HESA (2010) *Non-UK Domiciled Students* [Online] Available at: https://www.hesa.ac.uk/pr184 [Accessed 20 September 2014]

HESA (2011) *About HESA: Overview* [Online] Available at: http://www.hesa.ac.uk/index.php/content/view/4/54/ [Accessed 10 October 2011]

HESA (2012) *UKPIs: Non-Continuation Rates* [Online] Available at: https://www.hesa.ac.uk/pis/noncon [Accessed 1 August 2014]

Holman, D., Bass, J., Rouse, I. and Hobbs, M. (1999) "Population-based linkage of health records in Western Australia: development of a health services research linked database" *Australian and New Zealand Journal of Public Health* 23 pp.453 – 459

Homes Office (2012) *Immigration Statistics, July – September 2012* London: Home Office

Jaynes, E. "Confidence Intervals vs Bayesian Intervals" *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* 2 pp. 175-257

Jennissen, R. (2004) *Macro – economic Determinants of International Migration in Europe* Amsterdam: Dutch University Press

Judson, D. (2007) "Information integration for constructing social statistics: history, theory and ideas towards a research programme" *Journal of the Royal Statistical Society A* 170 pp. 483 – 501

Kim, K. & Cohen, J. (2010) "Determinants of international migration flows to and from industrialised countries: a panel data approach beyond gravity" *International Migration Review* 44 pp. 899 - 932

King, R. (2002) "Towards a new map of European Migration" *International Journal of Population Geography* 8 pp. 89 -106

King, R., Findlay, A. and Ahrens, J. (2010) *International Student Mobility Literature Review* Report to HEFCE, and co-funded by the British Council, UK National Agency for Erasmus, London: HEFCE

King, R. & Brooks, S. (2001) "On the Bayesian Analysis of population size" *Biometrika* 88 pp. 317 336

Knudsen, D. (1992) "Generalising Poisson regression: including a priori information using the method of offsets" *The Professional Geographer* 44 pp. 202 – 208

Kraly, E. & Gnanasekaran, K. (1987) "Efforts to Improve international migration statistics: A historical perspective" *International Migration* 21 pp. 967-995

Kritz, M., Lin, L. and Zlotnik, H. (eds) (1992) *International Migration Systems: A Global Approach* Oxford: Clarendon Press

Kupiszewska, D. Wisniowski, A. (2009) "Availability of statistical data on migration and migrant population and potential supplementary sources for data estimation" *Report for the MIMOSA Project*

Lathrop, R. (1967) "Perceived Variability" *Journal of Experimental Psychology* 73 pp. 498- 502

Lee, E. (1966) "A theory of migration" *Demography* 3 pp. 47 – 57

Liebig, T. and Sousa – Poza, A. (2002) "Migration, self-selection and income inequality: an international analysis" *Kyklos 5*7 pp. 125 – 146

Mason, A., Richardson, S. Plewis, I. and Best, N. (2011) "Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods" Working paper

Massey, D. (1988) "Economic Development and International Migration in Comparative Perspective" *Population and Development Review* 14 pp. 383 - 413

Massey, D. (1990) "The Social and Economic Origins of Migration" *Annals of the American Academy of Political and Social Science* 510 pp. 60 - 72

Massey, D. (2003) "Patterns and Processes of International Migration in the 21st Century" *Conference on African Migration in Comparative Perspective*

Massey, D., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, E. (1993) "Theories of international migration: a review and appraisal" *Population and Development Review* 19 pp. 431 - 466

Nelder, J. and Wedderburn, W. (1972) "Generalized Linear Models" *Journal of the Royal Statistical Society A* 135 pp. 370 – 384

McNees, S. (1990) "The role of judgement in macroeconomic forecasting accuracy" *International Journal of Forecasting* 6 pp. 287 - 299

O'Hagan, A. (1998) "Eliciting Experts Beliefs in Substantial Practical Applications" *The Statistician* 47 pp. 21-35

O'Hagan, A. (2005) *Uncertain Judgements: Eliciting Experts Probabilities* Chichester: Wiley

ONS (2001) *ONC Hard to Count Index* Titchfield: ONS [Online] Available at: http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-

conduct/the-one-number-census/methodology/steering-committee/key-papers/hard-to-count-index.pdf [Accessed 20 September 2014]

ONS (2003) *Discussion Paper: Proposals for an integrated population statistics system* Titchfield: Office for National Statistics

ONS (2006) *Update on ONS proposals for an integrated population statistics system* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/archive-material/components-of-the-imps-project/ipss/update-of-ipss.pdf [Accessed 20 September 2014]

ONS (2007) *A review of the potential use of administrative sources in the estimation of population statistics* London: UK National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/archive-material/archive-updates-and-reports/2007/a-review-of-the-potential-use-of-administrative-sources-in-the-estimation-of-population-statistics.pdf [Accessed 20 September 2014]

ONS (2010) *Travel Trends* Titchfield: Office for National Statistics [Online] Available at: www.ons.gov.uk/ons/rel/ott/travel-trends/2010/travel-trends---2010.pdf [Accessed 20 September 2014]

ONS (2011 a) *Long-Term International Migration Estimates, Methodology Document 1991 Onwards* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/rel/migration1/long-term-international-migration/2008/methodology-to-estimate-long-term-international-migration.pdf [Accessed 20 September 2014]

ONS (2011 b) "Long-term international migration estimates" *Information Paper: Quality and Methodology Information* Titchfield: Office for National Statistics

ONS (2011 c) *Improving estimates of international migration in Northern Ireland and between the UK and Republic of Ireland* Titchfield: Office for National Statistics

ONS (2011 d) *Improved Immigration Estimates to Local Authorities in England and Wales: Overview of Methodology* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/improvements-to-local-authority-immigration-estimates/overview-of-improved-methodology.pdf [Accessed 20 September 2014]

ONS (2012) *Migration Statistics Improvement Programme Final Report* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/msip-final-report/migration-statistics-improvement-programme-final-report---download-file.pdf [Accessed 20 September 2014]

ONS (2012 b) *Polish People in the UK - Half a million Polish Residents* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/dcp171780_229910.pdf [Accessed 20 September 2014]

ONS (2012 c) *Methods Used to Revise the National Population Estimates for Mid-2002 to Mid 2010* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/population-statistics-research-unit--psru-/methods-used-to-revise-the-national-population-estimates-for-mid-2002-to-mid-2010.pdf [Accessed 20 September 2014]

ONS (2012 d) *Explaining the Difference between the 2011 Census Estimates and the Rolled-Forward Population Estimates* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/population-statistics-research-unit--psru-/difference-between-the-2011-census-estimates-and-the-rolled-forward-population-estimates.pdf [Accessed 20 September 2014]

ONS (2013) *International Passenger Survey: Quality Information in Relation to Migration Flows, Background Note* Titchfield: Office for National Statistics [Online] Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/international-migration-methodology/international-passenger-survey-quality-information-in-relation-to-migration-flows.pdf [Accessed 20 September 2014]

Patel, J. & Read, C. (1996) *Handbook of the Normal Distribution* Cleveland OH :CRC Press

Peach, C. (2006) "South Asian migration and settlement in Great Britain, 1951 – 2001" *Contemporary South Asia* 15 pp. 133 - 146

Peck, J. (1996) *Work – Place: the Social Regulation of Labour Markets* New York: Guilford

Piore, M. (1979) *Birds of Passage: Migrant Labour and Industrial Societies* Cambridge: Cambridge University Press

Plummer, M., Best, N., Cowles, K., Vines, K (2006) "coda: Convergence diagnostics and output analysis for MCMC" *R News* 6 pp. 7-11

Pollard, N., Latorre, M. and Sriskandarajah, D. (2008) *Floodgates of Turnstiles. Post EU Enlgargement Migration Flows to (and from) the UK* London: IPPR

Portes, A. and Borocz, J. (1989) "Contemporary immigration: theoretical perspectives on its determinants and modes of incorporation" *International Migration Review* 23 pp. 606 - 630

Portes, A. and DeWind, J. (2004) "A cross-Atlantic dialogue: the progress of research and theory in the study of international migration" *International Migration Review* 38 pp. 828 – 851

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

Public Administration Select Committee (2013) *Migration Statistics* London: House of Commons

Putnam, R. (1993) "The prosperous community: social capital and public life" *American Prospect* 13 pp. 35 – 42

Ravenstein, EG. (1885) "Laws of Migration" *Journal of the Statistical Society of London* 48 pp. 167-235

Raymer, J. (2007) "Estimation of international migration flows: a general technique focused on the origin-destination association structure" *Environment and Planning A* 39 (4) pp. 985-995

Raymer, J., Abel, G. Disney, G. & Wisniowski, A. (2011 a) "Improving Estimates of Migration Flows to Eurostat" *CPC Working Paper 15* Southampton: ESRC Centre for Population Change

Raymer, J., Forster, J., Smith, P., Bijak, J., Wisnoiwski, A. (2011 b) "Integrated modelling of European migration: background specification and results" In: NORFACE (New Opportunities for Research Funding Agency Co-operation in Europe), *Integrated Modelling of European Migration, IMEM Workshop*. Chilworth, UK 25th – 27th May 2011

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

Raymer, J. and Bijak, J. (2009) "Modelling of statistical data on migration and migrant populations, MIMOSA" *Report on the technical consultancy in the United Kingdom*

Raymer, J., Rees, P., Blake, A., Boden, P., Brown, J., Disney, G., Lomax, N., Norman, P. and Stillwell, J. (2012) *Conceptual Framework for UK Population and Migration Statistics* Titchfield: ONS

Raymer, J. and Rogers, A. (2007) "Using age and spatial flow structures in the indirect estimation of migration streams" *Demography* 44 pp. 199 – 223

Raymer, J. and Smith, P. (2010) "Editorial: modelling migration flows" *Journal of the Royal Statistical Society A* 173 pp. 703 – 705

Raymer, J. Smith, P. and Guilietti, C. (2010) "Combining available migration data in England to study economic activity flows over time" *Journal of the Royal Statistical Society 4* 173 pp. 733 - 753

Raymer, J., Wisniowski, A., Forster, J., Smith, P. & Bijak, J. (2013) "Integrated Modelling of European Migration" *Journal of the American Statistical Association* 108 pp. 801 - 819

Rees, P. and Butt, F. (2004) "Ethnic change and diversity in England 1981-2001" *Area* 36: 174 – 186.

Rogers, A. and Castro L. (1981) "Age Patters of migration: cause specific profiles" in Rogers, A. (ed) *Advances in Multiregional Demography* Laxenburg Austria: International Institute for Applied Systems Analysis

Rogers, A., Little, J. and Raymer, J. (2010) *The Indirect Estimation of Migration: methods for dealing with irregular, inadequate and missing data* Dordrecht: Springer

Rogers, A., Raymer, J. and Willekens, F. (2002) "Capturing the age and spatial structures of migration" *Environment and Planning A* 34 pp. 341 -359

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

Rogers, A., Willekens, F. and Raymer, J. (2003) "Imposing age and spatial structures on inadequate migration-flow datasets" *The Professional Geographer* 55 pp. 56 – 69

Ryder, N. (1964) "Notes on a concept of population" *American Journal of Sociology* 69 pp. 447 – 463

Salt, J. (1989) "A comparative overview of international trends and types, 1950 – 1980" *International Migration Review* 23 pp. 431 – 456

Salt, J., Dobson, J., Koser, K., & McLaughlan, G. (2001) *International Migration and the United Kingdom: Recent Patterns and Trends*, RDS Occasional Paper 75, 2001, Home Office, London.

Salt, J. (2009) "International migration and the United Kingdom" *Report of the United Kingdom Sopemi to the OECD* London : University College London

Salt, J. and Millar, J. (2006) "International migration in interesting times: the case of the UK" *People and Place* 14 pp. 14 – 25

Sassen, S. (1991) *The Global City: New York, London, Tokyo* Princeton: Princeton University Press

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2011) *OpenBugs User Manual Version 3.2.1* [Online] Available at http://www.mrc-bsu.cam.ac.uk/bugs. [Accessed 6 May 2014]

Stagg, H., Jones, J., Bickler, G. & Abubakar, I. (2012) "Poor uptake of primary healthcare registration among recent entrants to the UK: a retrospective cohort study" *British Medical Journal: Open* 2 pp. 1-7

Stark, O. (2006) "Inequality and migration: a behavioural link" *Economic Letters* 91 pp. 146 – 152

Stark, O. and Taylor, E. (1989) "Relative deprivation and international migration" *Demography* 26 pp. 1 - 14

Statistics Commission (2004) "Census and Population Estimates and the 2001 Census in Westminster: Final Report" *Report No. 22* London: Statistics Commission [Online] Available at: http://www.statisticsauthority.gov.uk/reports---

correspondence/archive/statistics-commission-archive/research/report-22--census-and-population-estimates-and-the-2001-census-in-westminster-final-report.pdf [Accessed 20 September 2014]

Statistics New Zealand (2011) *Population Statistics domain plan 2011* Draft Report, Statistics New Zealand: Wellington

Stillwell, J. (1978) "Interzonal migration: some historical tests of spatial-interaction models" *Environment and Planning A* 10 pp 1187-1200.

Stillwell, J. (2005) "Inter-regional Migration Modelling: A Review and Assessment" In: *45th Congress of the European Regional Science Association,* Amsterdam: Vrije Universiteit

Stillwell, J. and Hussain, S. (2008) "Internal Migration of ethnic groups in England and Wales by Age and District Type" *Working Paper 08/3 School of Geography University of Leeds*

Stone, R. (1975) "An integrated system of demographic, manpower and social statistics and its links with the system of national economic accounts". *Sankhyā, The Indian Journal of Statistics*, 33(1 and 2) pp 1-184.

Stouffer, S, (1960) "Intervening Opportunities and Competing Migrants" *Journal of Regional Science* 2 pp. 1 – 26

Sumption, M. (2009) *Social Networks and Polish Immigration to the UK* London: IPPR

Taylor, E. (1986) "Differential migration, networks, information and risk" in Stark, O. (ed) *Research in Human Capital and Development* Vol 4 *Migration, Human Capital and Development* Greenwich: JAI Press

Todaro, M. (1969) "A Model of Labour Migration and Urban Unemployment in Less Developed Countries" *American Economic Review* 59 pp. 138 – 148

Townsend, P. (1985) "A sociological approach to the measurement of poverty – a rejoinder to Professor Amartya Sen" *Oxford Economic Papers, New Series* 4 pp. 659 - 668

Treasury Select Committed (2008) *Counting the Population: Eleventh Report of Session 2007-2008* London: House of Commons

UN (1998) *Recommendations on statistics of international migration*. New York: United Nations. [Online] Available at: http://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1e.pdf [Accessed 24 June 2013]

United Kingdom Statisitcs Authority (2013) *UK Statistics Authority Response to Public Administration Select Committee Report On Migration* London: UKSA [Online] Available at: http://www.parliament.uk/documents/commons-committees/public-administration/Letter-from-Sir-Andrew-Dilnot-061213.pdf [Accessed 20 September 2014]

Wallerstein, I. (1974) *The Modern World System, Capitalist Agriculture and the Origins of the European World Economy in the Sixteenth Century* New York: Academic Press

Wheldon, M., Raftery, A., Clark, S. & Gerland P. "Estimating demographic parameters with Uncertainty from fragmentary data" *Jounral of the American Statistical Association* 108 pp. 96 - 110

Willekens, F. (1983) "Log-linear modelling of spatial interaction" *Papers in Regional Science* 52 pp. 187 – 205

Willekens, F. (1994) "Monitoring international migration flows in Europe: towards a statistical data base, combining data from different sources" *European Journal of Population* 10 pp. 1-42

Willekens, F. (1999) "Modelling approaches to the indirect estimation of migration flows: from entropy to EM" *Mathematical Population Studies: An International Journal of Mathematical Demography* 7 pp. 239 – 278

Winkler, R. (1967) "The Assessment of Prior Distributions in Bayesian Analysis" *Journal of the American Statistical Association* 62 pp. 776-800

Wisniowski, A., Bijak, J., & Shang, HL. (2014) "Forecasting Scottish Migration in the Context of the 2014 Constitutional Change Debate" *Population Space and Place* 20 pp. 455-464

Wiśniowski, A., Bijak, J., Christiansen, S., Forster, JJ., Keilman, N., Raymer, J., Smith, P. (2013) "Utilising expert opinionto improve the measurement of international migration in Europe". *Journal of Official Statistics* 29 pp 583–607

Wisniowski, A., Keilman, N., Bijak, J., Solveig, C. Forster, J., Smith P., and Raymer, J. (2011) "Augmenting Migration Statistics with Expert Knowledge" *NORFACE MIGRATION Discussion Paper No. 2012-05*

Zlotnik, H. (1987) "The concept of international migration as reflected in data collection systems" *International Migration Review* 21 pp. 925 - 946