ESRC Centre for Population Change Working Paper

Number 15

Improving Estimates of Migration Flows to Eurostat

James Raymer

Guy J Abel

George Disney

Arkadiusz Wiśniowski

December 2011

ISSN2042-4116

ABSTRACT

In this paper we identify the current mandatory requirements and issues concerning the supply of detailed migration data to Eurostat. Using simple illustrations on immigration to the United Kingdom, we show how substantial and significant improvements can be made to the flows reported by the International Passenger Survey, which contain irregularities and missing data due to its relatively small sample size. Our general methodology is based on the idea of smoothing, repairing and combining data within multiplicative component framework.

KEYWORDS

Immigration; inadequate migration data; migration estimation; United Kingdom; Office for National Statistics; Eurostat.

EDITORIAL NOTE

Dr James Raymer, Dr Guy J Abel, George Disney and Arkadiusz Wisniowski are members of the ESRC Research Centre for Population Change, University of Southampton. Dr. Guy Abel is also at the *Wittgenstein Centre for Demography and Global Human Capital, Vienna Institute of Demography.*

Corresponding author: James Raymer, J.Raymer@soton.ac.uk

ACKNOWLEDGEMENTS

This report was commissioned by the Migration Statistics Unit within the Office for National Statistics (ONS) Centre for Demography**.** Raymer and Bijak visited the ONS in 2009 as part of their involvement in the Eurostat-funded project, entitled MIgration MOdelling for Statistical Analyses (MIMOSA). During this visit, they discussed the problems they encountered with UK's international migration data provided to Eurostat and presented some methodological options that ONS could explore for improving their international migration data. In this paper, we expand on these ideas, as well as those contained in the recent book by Rogers et al. (2010) on smoothing, repairing and inferring inadequate or missing migration data.

ESRC Centre for Population Change

The ESRC Centre for Population Change (CPC) is a joint initiative between the Universities of Southampton, St Andrews, Edinburgh, Stirling, Strathclyde, in partnership with the Office for National Statistics (ONS) and the National Records of Scotland (NRS). The Centre is funded by the Economic and Social Research Council (ESRC) grant number RES-625-28-0001.

Website | Email | Twitter | Facebook | Mendeley

IMPROVING ESTIMATES OF MIGRATION FLOWS TO EUROSTAT

TABLE OF CONTENTS

1. INTRODUCTION

This paper details work commissioned by the Migration Statistics Unit within the Office for National Statistics Centre for Demography (ONSCD). The aim of this work is to deliver a recommendation regarding how the Office for National Statistics (ONS) could improve the quality of detailed estimates of migration flows required by Eurostat, to include methodology, and estimates of the quality improvement that would be achieved. In response to this aim, we first identify the current mandatory requirements and issues concerning the supply of migration data to Eurostat. We then introduce several estimation techniques and strategies that can be used to overcome these obstacles.

Our strategy for improving the IPS data includes three methodological options (Rogers et al. 2010). The first involves smoothing the data. We use the term “smoothing” to represent the process of limiting the effect of randomness on the age, spatial or temporal patterns of migration caused by natural variation or variation due to insufficient sample size. This may involve (i) fitting a line or curve to a particular pattern of migration or (ii) removing higher-order interaction effects in a log-linear model for a contingency table of migration flows. The second relies on “imposing” methods, which borrow age or spatial patterns of migration from other patterns, e.g., when an average age profile of immigration is used to represent the age profile of immigration from a small country not captured adequately in the reported data. The third methodological option involves “inferring” migration, which borrows age and / or spatial data from auxiliary sources that serve as useful proxies for the particular migration pattern that requires estimation.

2. EUROSTAT’S REQUIREMENTS FOR REPORTING INTERNATIONAL MIGRATION FLOWS

In this section, we outline Eurostat’s mandatory requirements for immigration and emigration and briefly describe ONS’s current method for producing international migration statistics based on the International Passenger Survey (IPS), asylum seeker and refugee data from the Home Office and flows between Ireland and the UK provided by Ireland.

2.1. EUROSTAT'S REQUIREMENTS

The following information is taken from Article 3 of the European Parliament Regulation (EC) No. 862/2007.[1] 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.

In addition to these requirements, member countries are encouraged to supply other migration data, such as immigration flows by country of previous residence, on a voluntary basis. The complete list of mandatory and voluntary requirements from Eurostat are summarised in Table 1. The Eurostat names for the tables are also included. Refer to the Appendix for the matching of countries to country groups, which are defined as follows:

EU27 – 27 member states of the European Union
EFTA – The European Free Trade Association
CC3_07 – European Union Candidate Countries
HDC – Non-EU Highly Developed Country
MDC – Non-EU Medium Developed Country
LDC – Non-EU Low Developed Country

[1] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32007R0862:EN:NOT

Name of Table	Mandatory Requirements	Voluntary Requirements
IMM1CTZ Immigrants by citizenship, sex, age group	Citizenship by groups of countries, sex, 5 year age groups	Citizenship by individual countries, sex, 5 year age groups
IMM2CTZ Immigrants by single year of age: nationals and non-nationals	Citizenship by foreigners/ nationals/unknown, sex, single year of age	NA
IMM3CTB Immigrants by country of birth, sex, age group	Country of birth by groups of countries, sex, 5 year age groups	Country of birth by individual countries, sex, 5 year age groups
IMM4CTB Immigrants by single year of age: native born and foreign born	Country of birth by foreigners/nationals/unknown, sex, single year of age	NA
IMM5PRV Immigrants by country of previous residence, sex, age group	Country of previous residence by groups of countries, sex, 5 year age groups	Country of previous residence by individual countries, sex, 5 year age groups
EMI1CTZ Emigrants by citizenship, sex, age group	Total; total by sex; total by groups of countries; total by 5 year age group	Citizenship by groups, individual countries, sex, 5 year age groups
EMI2 Emigrants by sex and single year of age	Total by single year of age; total by sex	By sex and single year of age
EMI3NXT Emigrants by country of next usual residence, sex, age group	Total; total by sex; total by 5 year age group; totals by EU, non-EU and unknown	Country of next residence by sex and five year age group

Table 1. Eurostat's mandatory and voluntary data requirements for international migration flow data

2.2 THE ONS METHOD FOR ESTIMATING TOTAL INTERNATIONAL MIGRATION

There is no single source of data that captures all long-term international migration to and from the United Kingdom. As a result, ONS uses a combination of data from different sources. Each source of data has different characteristics that can be used to help estimate international migration. However, it is important to note that none of the data sources used are designed specifically to measure international migration. The current estimates of Long Term International Migration (LTIM) are comprised from the following estimated components: International Passenger Survey, Northern Ireland flows, visitor switchers, asylum seekers and migrant switchers.

The following information on the ONS method for estimating for long-term international migration was taken from a recent ONS document entitled "Long-Term International Migration Estimates, Methodology Document, 1991 onwards."[2] ONS applies the United Nations recommended definition of an international long-term migrant. That is, a long-term international migrant is defined "as someone who changes his or her country of usual residence for a period of at least a year, so that the country of destination effectively becomes the country of usual residence." This definition of international migration forms the conceptual basis of the question design of the international migration section of the International Passenger Survey (IPS) (Boden and Rees 2010).

International Passenger Survey: Passengers are asked about their intentions, to determine whether they intend to stay in the UK upon arrival, or in their destination upon departure, for at least 12 months. As a result, the figures for immigration and emigration obtained from the IPS represent intentions and not actual length of stay. As reported in the ONS documentation, the IPS has several limitations with regard to measuring immigration and emigration. First, it is a sample survey and therefore only a small fraction of migrants from and to the UK are captured. Second, it does not capture asylum seekers who may be entering or leaving the UK, or migrants between the UK and the Republic of Ireland. Finally, it does not take into account the changing intentions of passengers.

[2] Available at: www.statistics.gov.uk/downloads/.../Methodology-to-estimate-LTIM.pdf

The IPS is a multi-purpose sample survey of passengers arriving at, and departing from, the United Kingdom’s air and sea ports and the Channel Tunnel. In 2007, the IPS sample was over 300,000 and had an overall response rate of 80 percent. About 1.5 percent of those sampled were migrant interviews, which amounted to 4,450 persons. The IPS sample is stratified to ensure that it is representative by mode of travel, route and time of day. Interviews are conducted throughout the year. The information collected by the survey is weighted to produce national estimates of immigration and emigration, including breakdowns by country of origin/destination, citizenship, age and sex.

For 2007, the overall standard error for the estimated total immigration of 527,000 migrants was 3.8 per cent. This gives a range of between 488,000 and 566,000 as the 95 per cent confidence interval for the IPS estimate of the number of migrants entering the UK during 2007 (obtained as +/- 1.96 times the standard error). For the 2007 emigration flow of 318,000 migrants, the standard error was 4.3 per cent. This gives a range of 291,000 to 345,000 migrants as the corresponding 95 per cent confidence interval. When estimates are broken down into further detail, greater care must be taken with their interpretation. This is because these estimates will be based on a smaller number of survey contacts, which increase the uncertainty around the estimate. For example, it is not possible to produce estimates for a single year for most individual citizenships or countries of last/next residence because of the small number of survey contacts that comprise each estimate.

As mentioned previously, a key feature of the IPS question design is that it is based on passenger intentions. The ONS has developed methods that take into account migrants whose intentions, with regard to length of stay, change. This group of people are known as switchers. There are two types of switchers. Firstly, those whose intention it is to enter or leave the UK as a visitor (i.e., a stay of less than 12 months) but actually end up staying for more than 12 months. These visitors who become migrants are known as “visitor switchers.” Secondly, those whose intention it is to enter or leave the UK as a migrant (i.e., a stay of more than 12 months) but actually end up staying or leaving for less than 12 months. These migrants who become visitors are known as “migrant switchers”. Both types of switchers are estimated.

Asylum seekers: The Home Office is responsible for immigration control. They provide data for different types of asylum seekers: applications, refusals, appeals, returnees and application withdrawals. This information is used to identify the number of asylum seekers who qualify under the definition of a long-term international migrant and are used as part of the Total International Migration (TIM) estimates.

Republic of Ireland: Until 2007, data from the Central Statistics Office (CSO) in Ireland were used to estimate the flows between Ireland and the UK. This was necessary because the IPS did not survey any of the routes between Ireland and the UK until 1999. However, when IPS flow estimates were compared to the estimates from the CSO it was concluded that the CSO was underestimating migration flows between the UK and Ireland. As such the ONS, since 2008, has used the IPS to estimate migration between the UK and Ireland.

Northern Ireland: Until 2007, the IPS was used to estimate migration to and from Northern Ireland. However, there were concerns about the reliability of these estimates, mainly because the IPS did not survey any of the ports in Northern Ireland. Therefore, from 2008 onwards, the ONS incorporated Northern Ireland’s Statistics and Research Agency’s (NISRA) estimations of long term international migration into their TIM estimate. NISRA use health card data to identify international migrants for their population estimates. A limitation of using this method is that it does not account for short term migrants and switchers; however, the benefit of having a more reliable account of international migration to and from Northern Ireland is thought to outweigh these limitations.

3. ASSESSMENT OF IMMIGRATION FLOW DATA PROVIDED BY IPS

Migration data from the International Passenger Survey (IPS) are assessed in relation to Eurostat’s requirements. For illustration, we focus on the tables (Immigrants by citizenship, sex and age group IMM1CTZ and immigrants by country of previous residence, sex and age group IMM5PRV) to identify the relative strengths and weaknesses of the IPS data. As the IPS captures approximately 90% of the flows, and is thus the most important source of data, it represents the main focus of this section and remainder of this paper.

The main issue concerning the United Kingdom's supply of international migration flow data to Eurostat is that the primary source of data are based on a passenger survey, which does not contain large enough sample sizes to meet the required level of detail. For many of the requirements, the survey estimates result in data of very poor or unacceptable quality. In fact, Raymer and Bijak (2009) stated that "…the migration flow data provided to Eurostat in recent years have been of such poor quality that they have been deemed unusable for understanding changes in the spatial and age patterns over time."

In this section, we show how the IPS data appear at various levels of disaggregation. As the levels of disaggregation increase, we expect the relative quality of data to decrease. While it can be difficult to distinguish between actual patterns and sample fluctuations, the aim of this analysis is to identify where the data are likely to become unreliable. In general, we expect the patterns to be stable over time, particularly for large or established flows.

According to the IPS data, immigration to the UK increased from 350 thousand in 2000 to around 500 thousand or more from 2004 onwards (see Figure 1). The reason for the large jump in the number of migrants in 2004 was due to the European Union adding 10 new countries (with substantially lower per capita GDP than other members of the EU) to its membership in 2004, for which migrants from these countries obtained immediate access and employment rights in the UK.

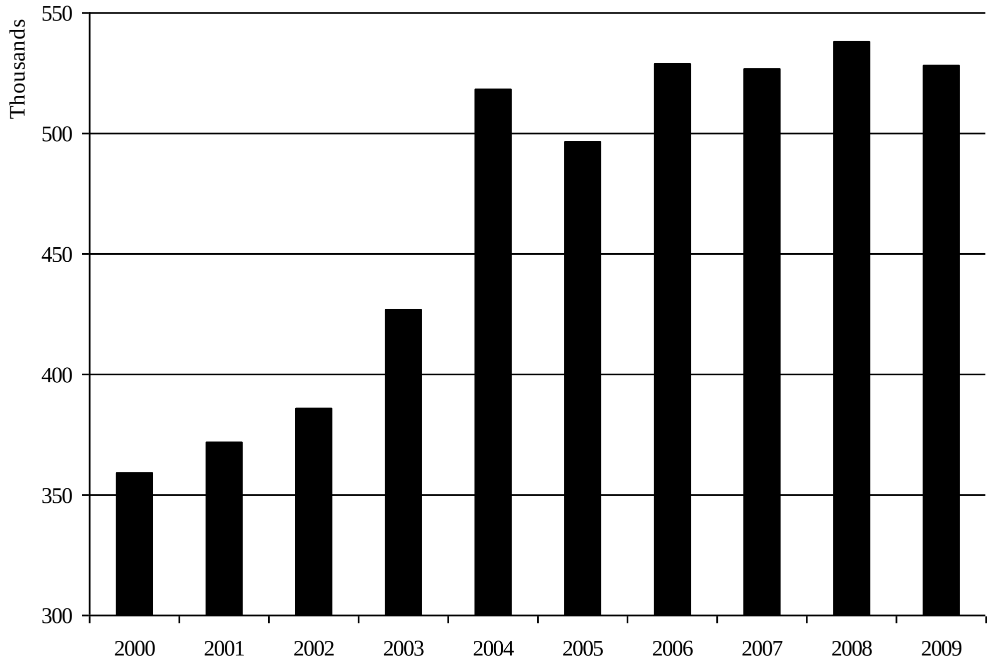


Figure 1. Total immigration to the United Kingdom, 2000-2009

The proportions of total immigration by age are shown in Figure 2 for the years 2000 to 2009. Here, we find strong regularities in the patterns over time with some minor fluctuations in the child, young adult and age 45+ age groups. The total flows by sex presented in Figure 3, on the other hand, show a remarkable divergence in 2005 and onwards, where the female flows become substantially lower than male flows. We cannot think of a logical reason for this. It could be due to the recent influx of EU accession migrants or other changes in the patterns. It could also be due to a coding or sampling issue with the IPS. For modelling purposes, we would like to assume that the overall age and sex structures in the IPS data are reliable. Finally, the age and sex patterns of total immigration are presented in Figure 4. Here, the age and sex patterns are largely stable over time, which is good for the purpose of estimation. The male age profiles exhibit a wider labour force peak than do the females.

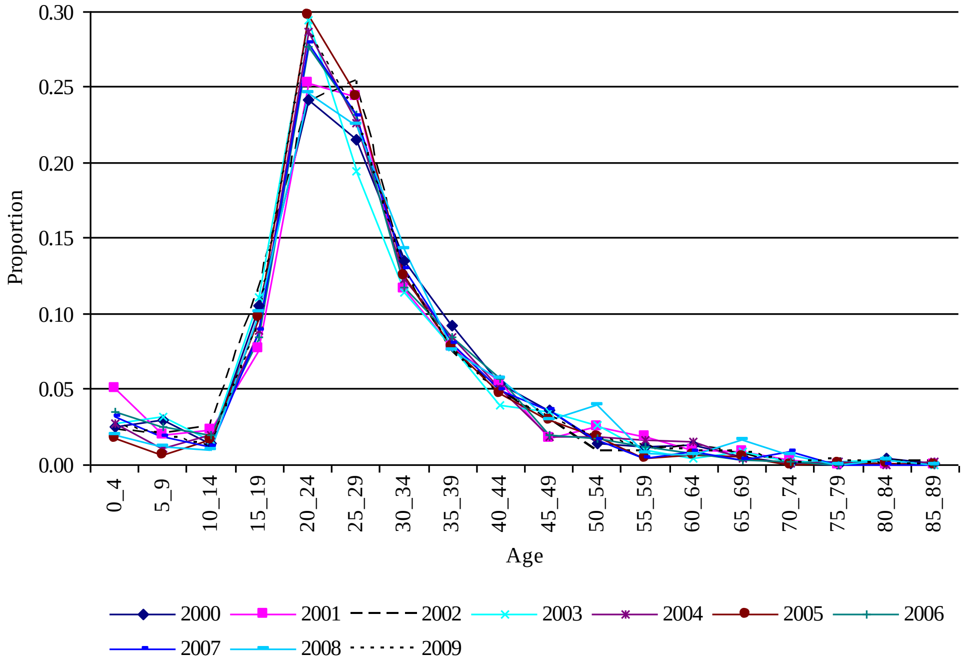


Figure 2. Immigration to the United Kingdom by age, 2000-2009

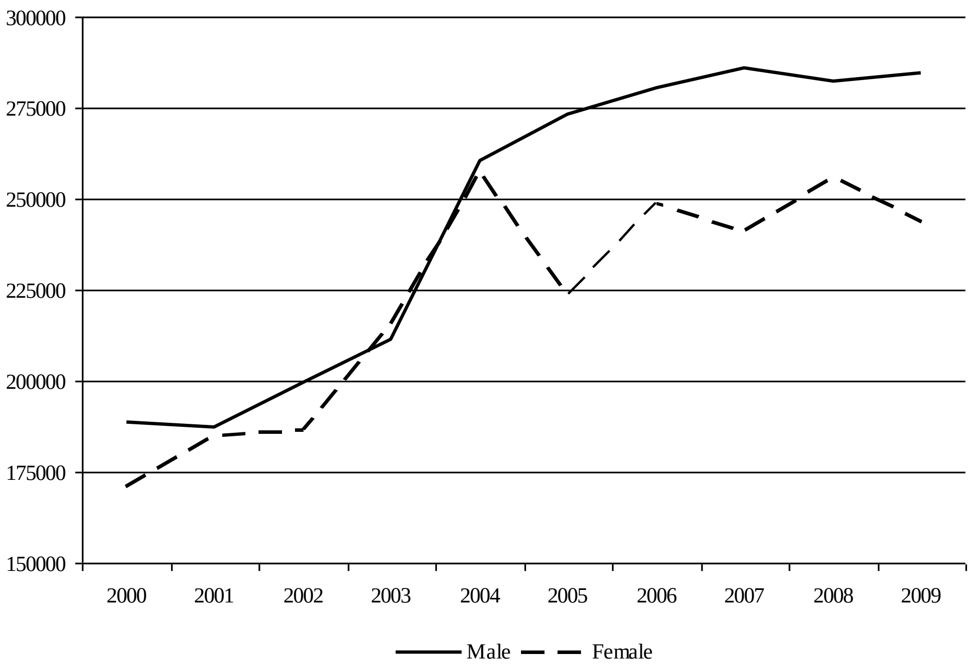


Figure 3. Immigration to the United Kingdom by sex, 2000-2009

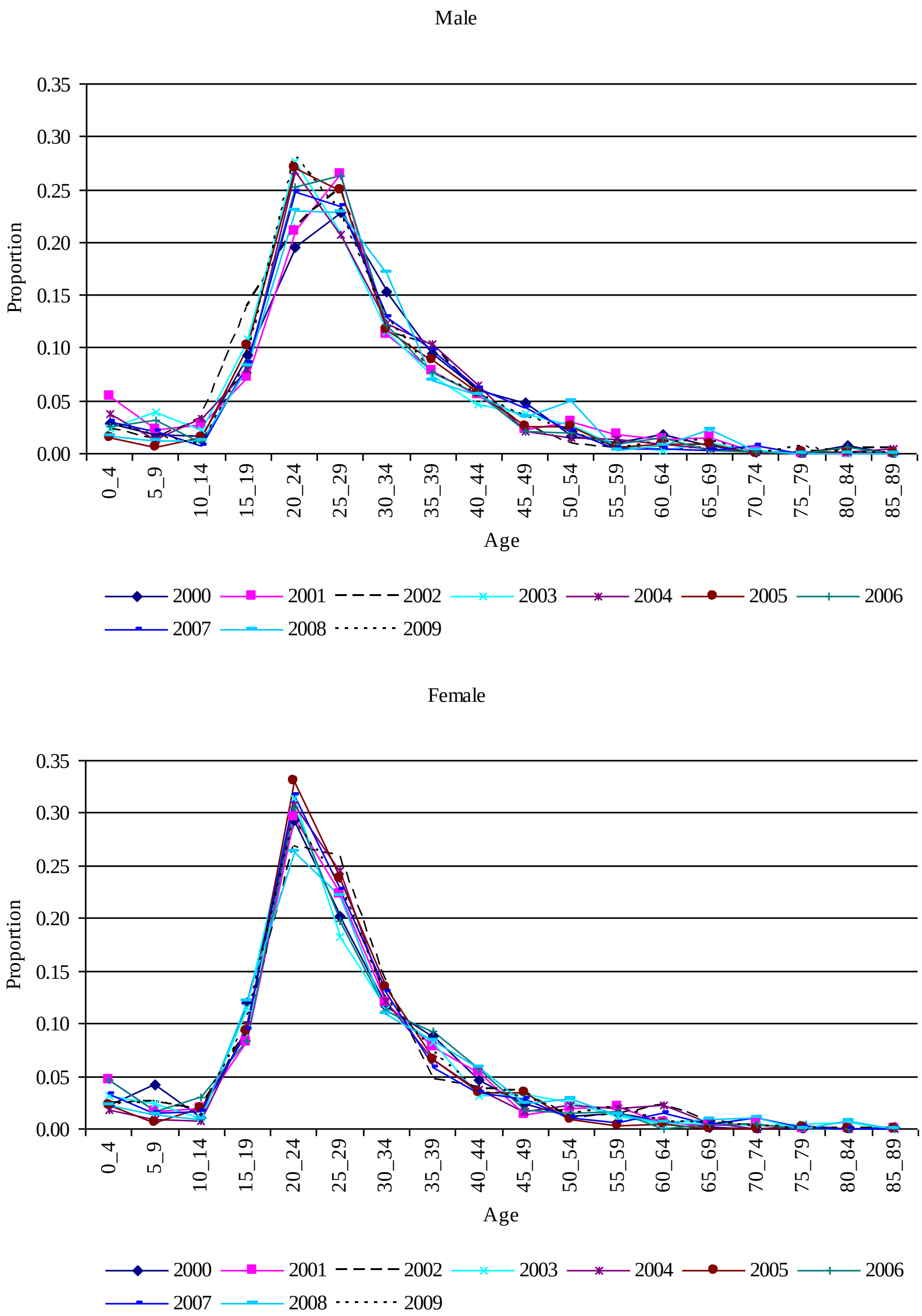


Figure 4. Immigration to the United Kingdom by age and sex, 2000-2009

Based on the analysis of the patterns above, we believe that the overall age and sex patterns of immigration to the UK revealed in the IPS are reasonable and reliable, with the possible exception of the overall sex patterns. In the next two subsections, the age-specific flows are disaggregated by citizenship group and country of previous residence, respectively.

3.1 IMMIGRATION BY AGE, SEX AND CITIZENSHIP GROUP

Eurostat requires seven groups to be identified in the citizenship flow tables. These include future accession countries (CC3 07), countries in the EFTA, nationals (United Kingdom), current EU countries (EU27), High Developed Countries (HDC), Low Developed Countries (LDC) and Medium Developed Countries (MDC). The immigration flows by citizenship group are presented in Figure 5.

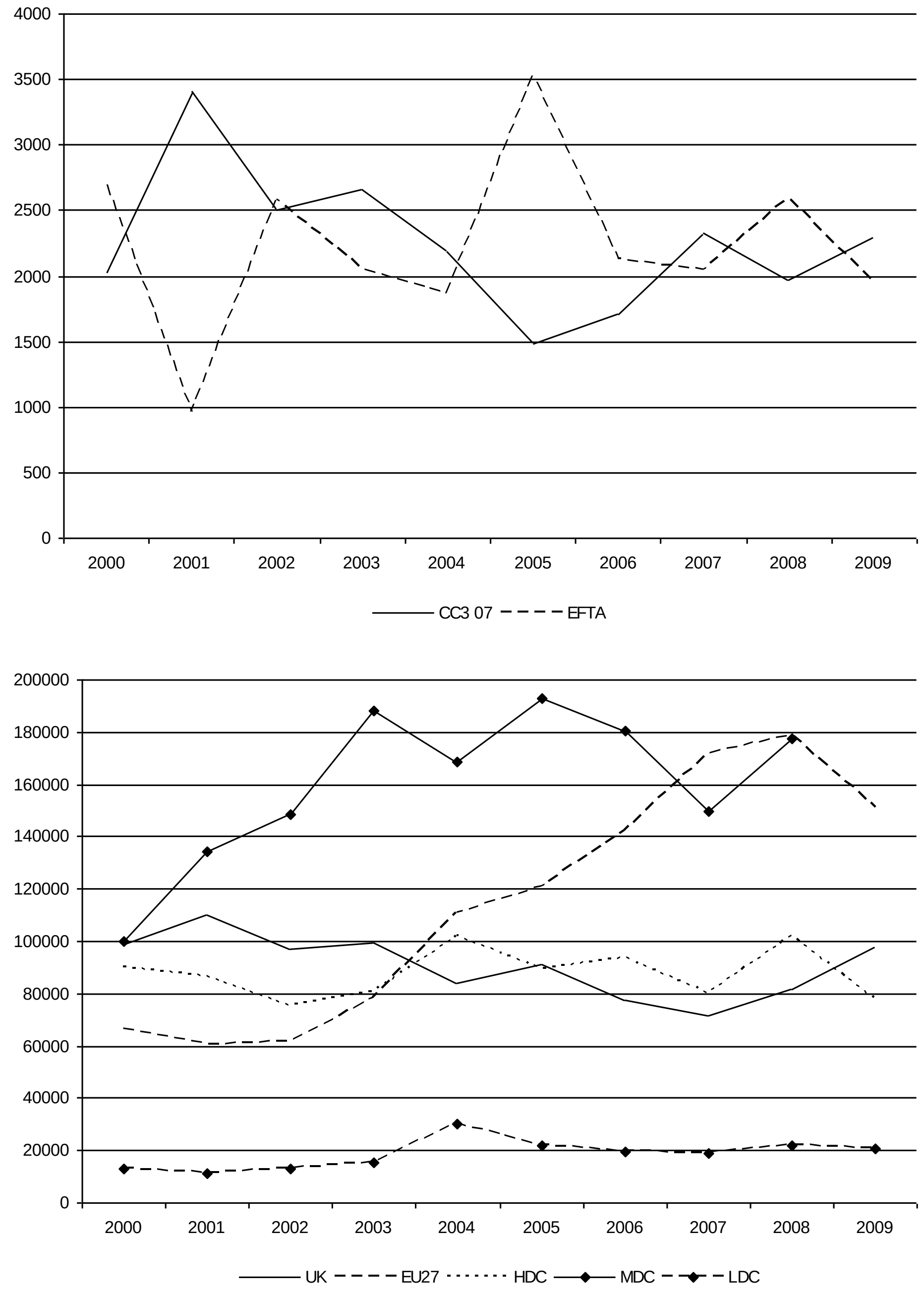


Figure 5. Immigration to the United Kingdom by citizenship group, 2000-2009

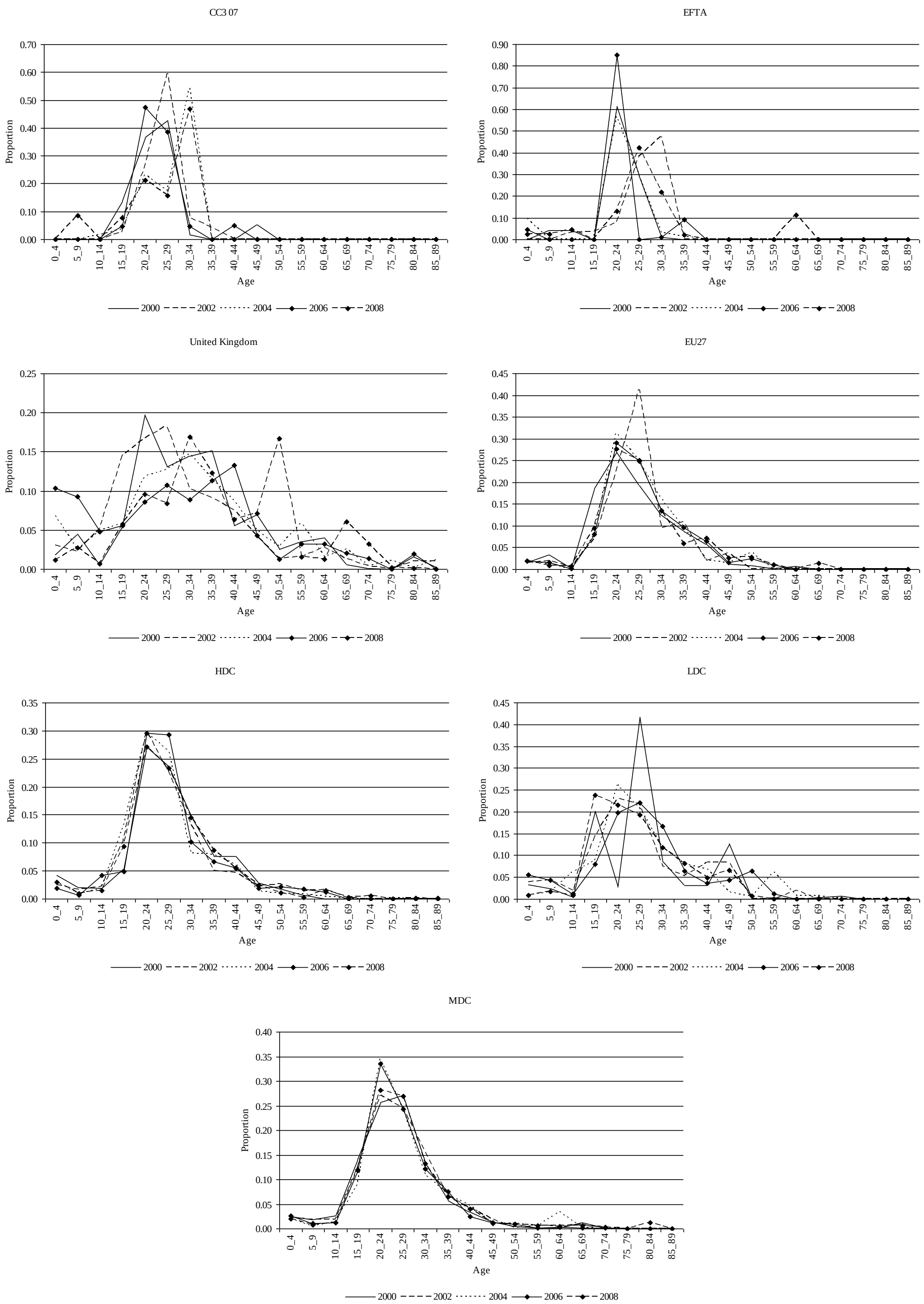


Figure 6. Immigration to the United Kingdom by age and citizenship group, 2000-2009

Note: CC3_07 European Union Candidate Countries, EFTA The European Free Trade Association, EU27 27 member states of the European Union, HDC Non-EU Highly Developed Country, MDC Non-EU Medium Developed Country, LDC Non-EU Low Developed Country.

The corresponding age-specific proportions of these seven groups are presented in Figure 6. Clearly, the IPS struggles to capture the patterns of the two smaller groups consisting of CC3 07 and EFTA migrants with average flows of just over two thousand per year. Also, the LDC group, with an average flow of 19 thousand, is fairly irregular. The smoothest age profiles appear for the HDC and MDC migrants with average flows of 88 thousand and 153 thousand, respectively, and to some extent the EU27 migrants with an average flow of 114 thousand. The reason why the age patterns of UK nationals are so irregular, with an average flow of 91 thousand, is not clear. Based on the sizes of these flows, they should appear more regular.

To further illustrate the problems with the sample size in the IPS data, consider the plots in Figure 7, which includes the proportion of the total citizenship group flows that are males from 2000 to 2009. Here, we see that percent males in the EFTA flows vary from around 10 percent to 85 percent, depending on the year. The flows for the larger citizenship groups are more stable over time, varying from around 40 percent to 65 percent.

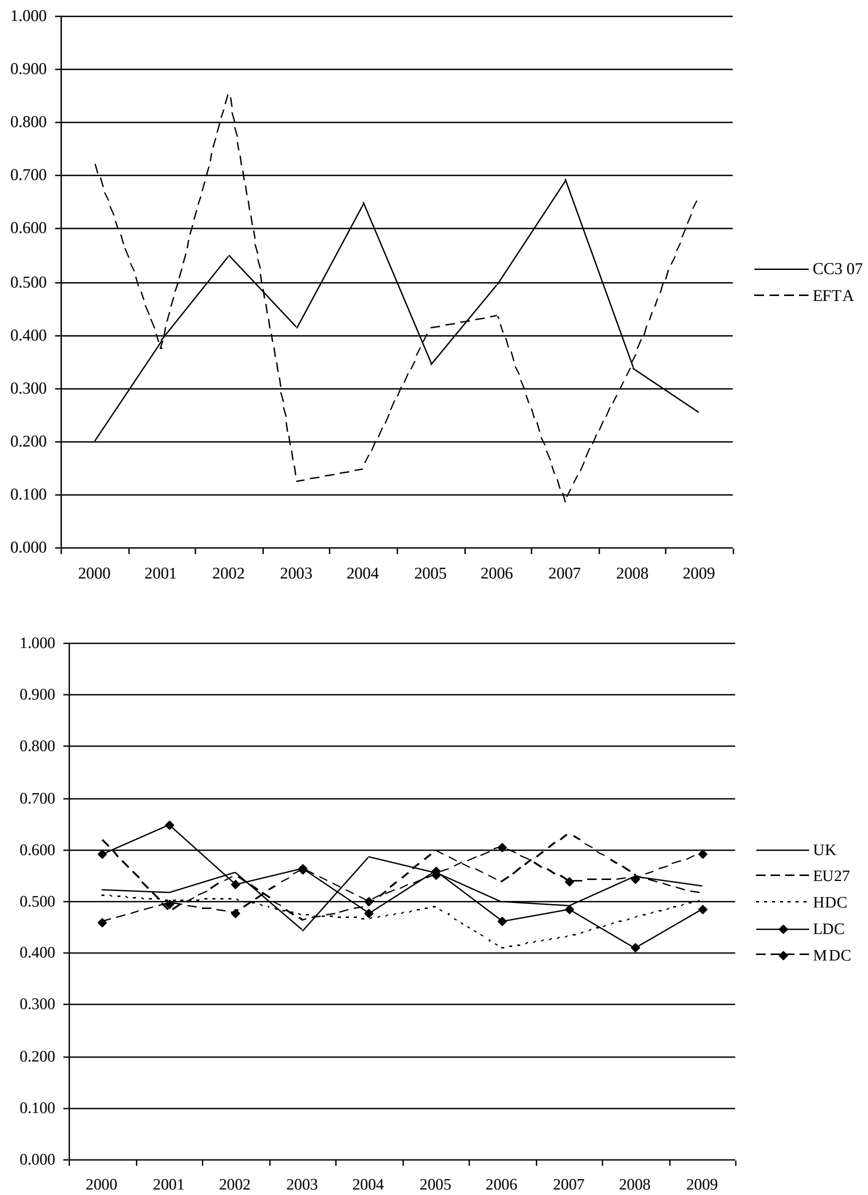


Figure 7. Proportion males in the immigration to the United Kingdom flows by citizenship group, 2000-2009

3.2 IMMIGRATION BY AGE, SEX AND COUNTRY OF PREVIOUS RESIDENCE

For the immigration flows by age, sex and country of previous residence, the same problems we found in the previous subsection appear again. The flows by country group of previous residence are shown in Figure 8. The EU27, HDC, MDC exhibit the most stable patterns, followed by LDC. The CC3 07 and EFTA flows are clearly not reliable.

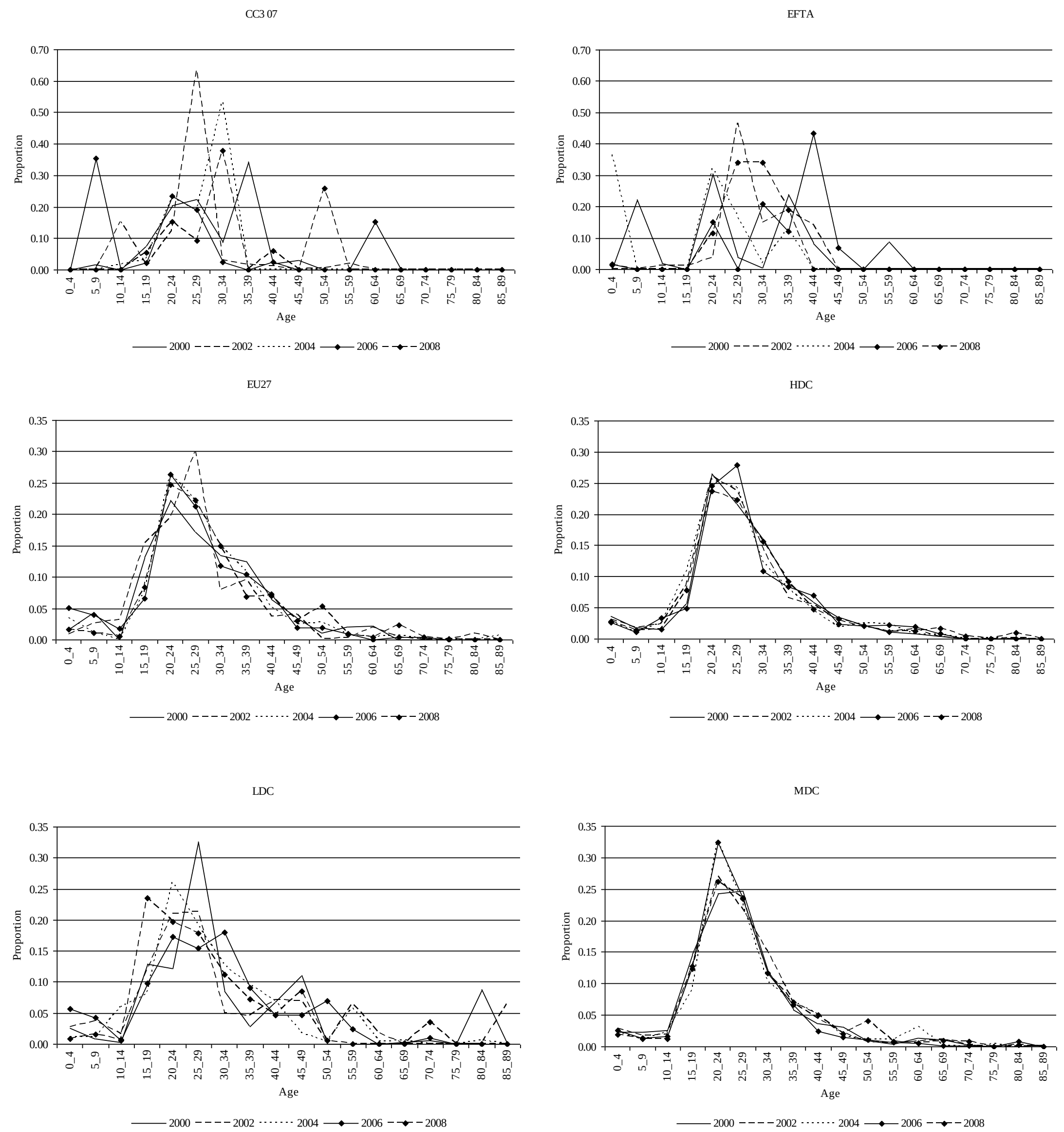


Figure 8. Immigration to the United Kingdom by age and country group of previous residence, 2000-2009

Note: CC3_07 European Union Candidate Countries, EFTA The European Free Trade Association, EU27 27 member states of the European Union, HDC Non-EU Highly Developed Country, MDC Non-EU Medium Developed Country, LDC Non-EU Low Developed Country.

Since we know that larger flows result in more reliable patterns, we next show how this relates to country-specific immigration flows. In Table 2, we show the top senders of migrants to the UK in terms of their totals summed from 2000 to 2009. According to the IPS, India sent 390 thousand migrants over the ten year period, followed by Australia with 360 thousand, Poland with 308 thousand, China with 300 thousand and the United States of America with 254 thousand. These five flows are shown for each year in the top panel of Figure 9. The lower plot contains flows from five countries sending between 70 thousand and 110 thousand migrants over the ten year period. Here, we see that there is considerably more year-to-year variability exhibited by countries sending 70-110 thousand than for the larger sending countries presented in the upper plot. Finally, a selection of age-specific flows for the top senders is presented in Figure 10. While some flows appear reasonable (e.g., Australia, China and India), most contain unexpected irregularities across age groups.

	Group	Country	Total	Average
1	MDC	India	390,484	39,048
2	HDC	Australia	359,601	35,960
3	EU27	Poland	307,832	30,783
4	MDC	China (exc. Taiwan)	300,015	30,001
5	HDC	United States of America (USA)	253,729	25,373
6	MDC	South Africa	222,401	22,240
7	MDC	Pakistan	188,991	18,899
8	EU27	Ireland (2008-9)	26,807	13,403
9	EU27	Spain	126,302	12,630
10	HDC	New Zealand	125,407	12,541
11	MDC	Philippines	108,431	10,843
12	HDC	Canada	77,517	7,752
13	LDC	Nigeria	75,260	7,526
14	MDC	Bangladesh	71,537	7,154
15	HDC	Japan	70,165	7,016
16	HDC	Malaysia	69,387	6,939
17	EU27	Netherlands	67,733	6,773
18	EU27	Italy	66,771	6,677
19	LDC	Zimbabwe	48,187	4,819

Table 2. Top senders of immigration to the United Kingdom according to the International Passenger Survey, 2000-2009

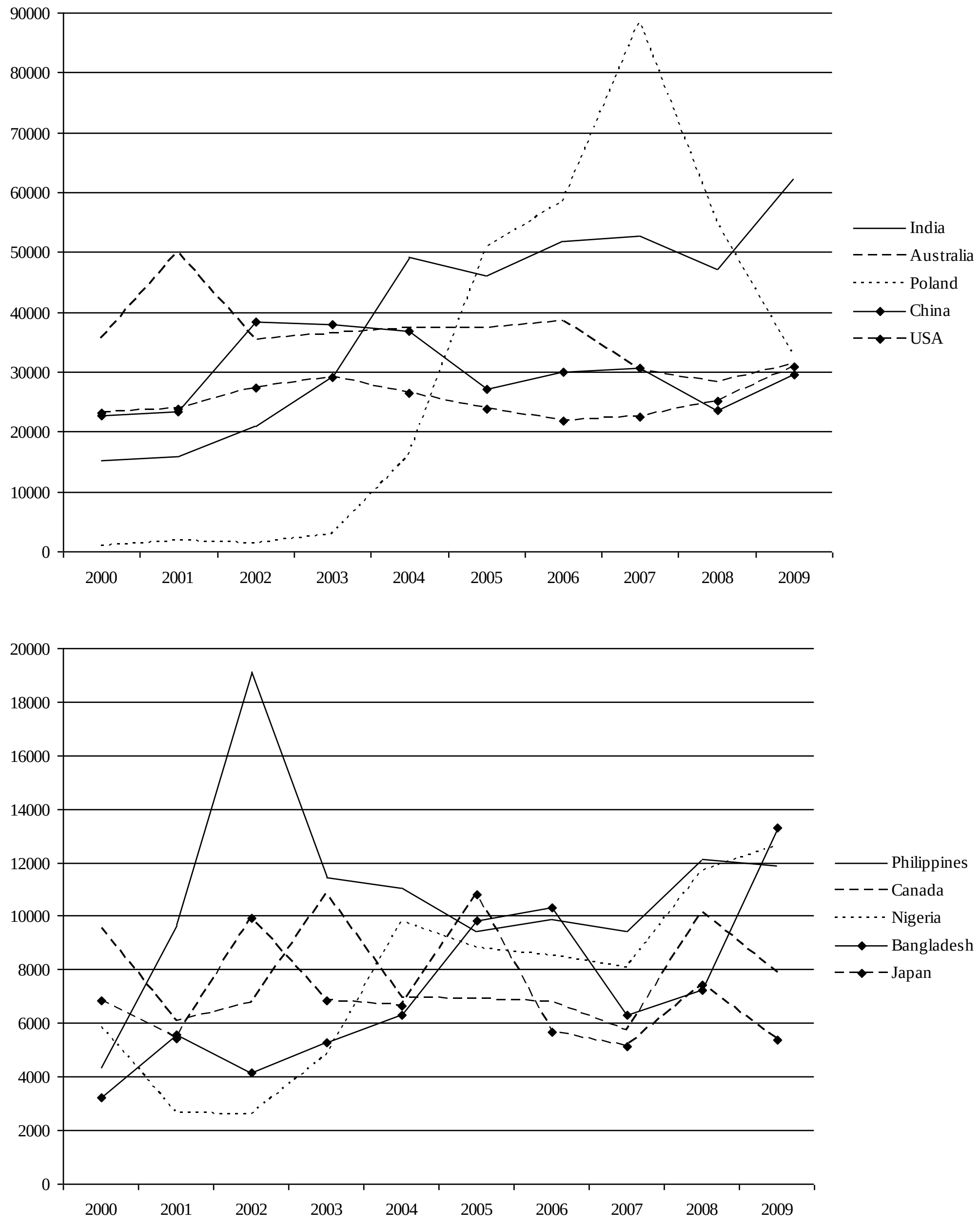
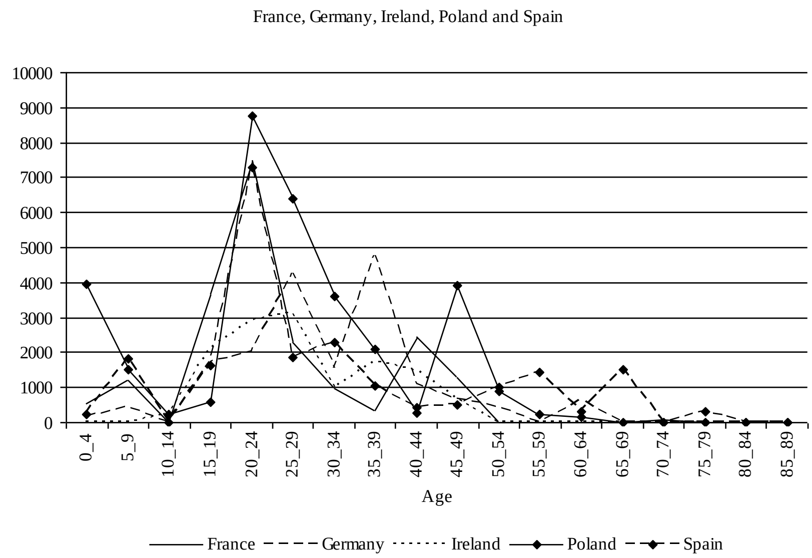


Figure 9. Immigration to the United Kingdom by selected countries of previous residence, 2000-2009: Countries with average flows greater than 25,000 and countries with average flows between 7,000 and 11,000 per year



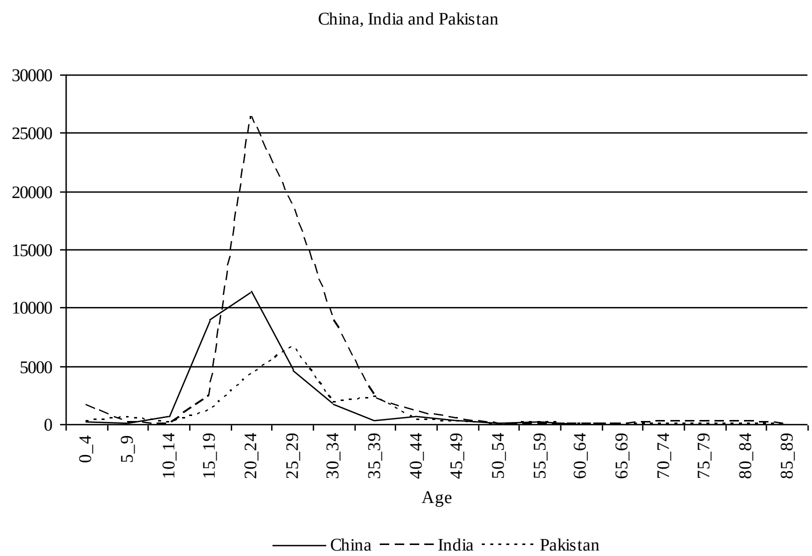


Figure 10. Immigration to the United Kingdom by age and selected countries of previous residence, 2009

3.3 SUMMARY

In this section, we have shown how irregularities across age, sex and country groups appear for flows by citizenship and country of previous residence. In the next three sections, we introduce methods for smoothing, repairing and inferring migration patterns, respectively. The data presented in this section is used as the basis for illustration the three estimation approaches.

4. SMOOTHING METHODS FOR IMPROVING IPS DATA

In this section, we present three methods that can be used to smooth the data: pooling data, fitting model migration schedules and unsaturated log-linear modelling. We focus mostly on age patterns, although the ideas and methods can be extended to other variables in the data.

4.1 POOLING DATA

The method of pooling can be used to smooth the data by averaging patterns over time. For illustration, consider the data presented in the left-hand side of Figure 8: immigration by age from CC3 07, EU27 and LDC countries. For this exercise, we first estimate the total levels of immigration based on three-year moving averages. Second, we assume the aggregate totals by country group are accurate and smooth only the age profiles according to a 10-year average and 3-year rolling averages. In this latter case, the averaged age profiles are rescaled to match the total level of migration for each year. The results for CC3 07, EU27 and LDC total immigration flows from 2001-2008 are presented in Figure 11. The age specific flows for the same groups are presented in Figure 12. We find that pooling is useful for reducing the variation in all flows, however, with less success for relatively small groups.

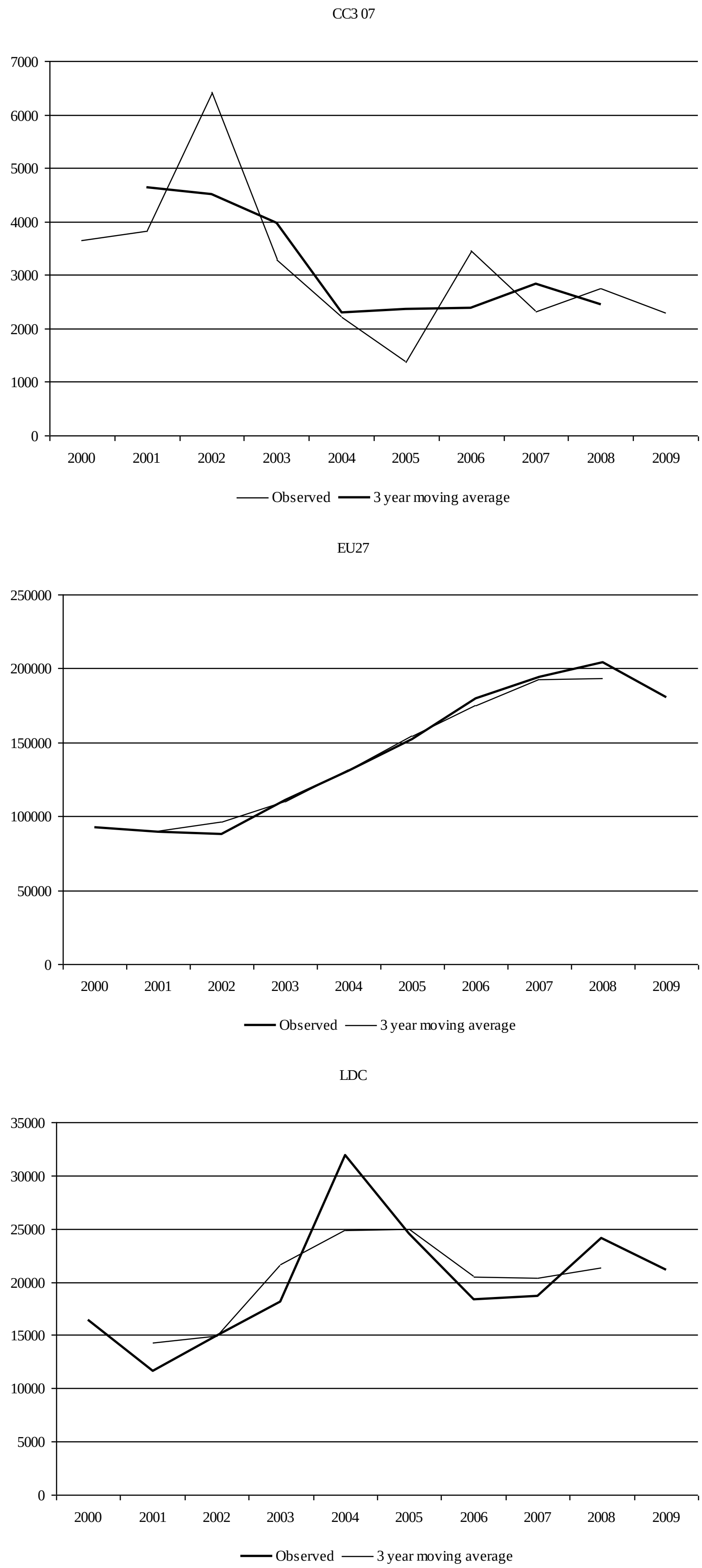
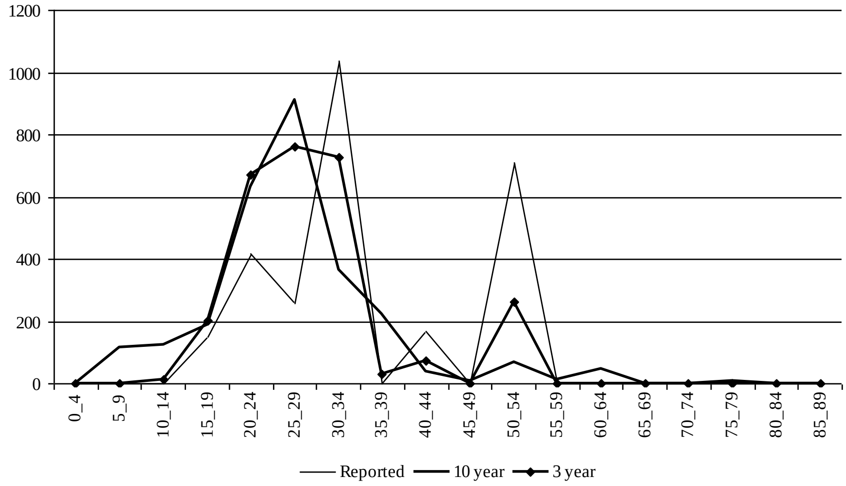
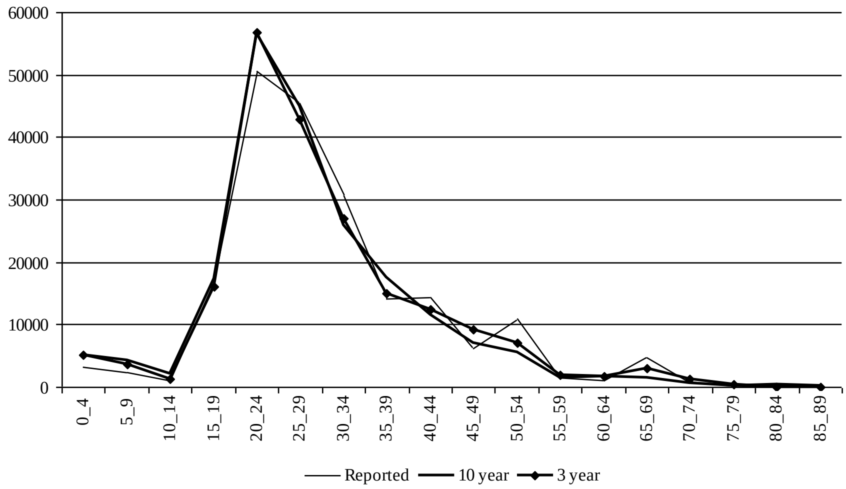


Figure 11. Reported and predicted (3-year moving average) immigration from CC3 07, EU27 and LDC countries, 2001-2008

CC3 07 2008



EU27 2008



LDC 2008

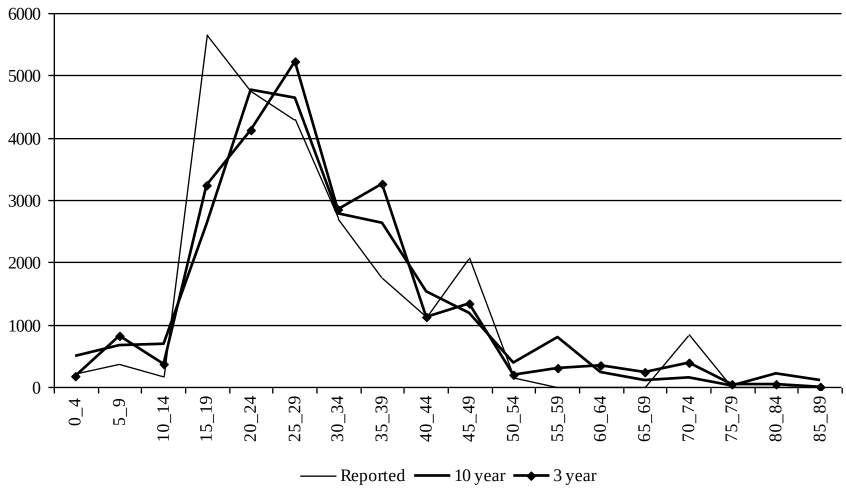


Figure 12. Reported and predicted (3-year moving average) age-specific immigration from CC3 07, EU27 and LDC countries, 2008

4.2 FITTING MODEL SCHEDULES TO AGE PATTERNS

Linear and non-linear regression lines can be fitted to IPS data for the purposes of smoothing. In this subsection, we focus on the more complicated non-linear regression models designed for age-specific migration.

Migration propensities differ greatly according to age. Typically, an age-specific profile of migration shows a downward slope from the early childhood age groups to about age sixteen, is followed by a rise to a peak in the young adult age groups (usually around age twenty-two), then gradually tapers off to the oldest age groups. This "standard" age profile of migration can be fully described using a multiexponential model migration schedule (Rogers and Castro 1981; Rogers and Little 1994; Rogers et al. 2010). While there are several variants of model migration schedules, the one most often used is the seven parameter version:

$$N_{ix} = a_0 + a_1 \exp(-\alpha_1 x) + a_2 \exp\{-\alpha_2(x-\mu_2) - \exp[-\lambda_2(x-\mu_2)]\}, \quad (1)$$

where N_{ix} denotes standardized (to unit area) age profiles of migration from, say, country i at age group x. The a_0, a_1, and a_2 are level parameters, whereas the α_1, α_2, μ_2, and λ_2 parameters are shape parameters.

For illustration, model migration schedules were fitted to the EU27 and LDC data presented in Figure 8. These data represent cases where the data are in need of smoothing. Applying model migration schedules to smooth the corresponding CC3 07 and EFTA data would not be appropriate as they do not exhibit any sort of migration age profile that we expect. Methods to deal with these country groups are discussed in Sections 5 and 6.

To fit model migration schedules to the observed IPS data, we used the statistical package TableCurve2D, which has a very useful graphical interface. However, these models can be fitted by non-linear regression routines found in most standard statistical packages, such as Stata, SPSS or SAS. To get these models to fit, it is important to have reasonable starting parameter values, which makes the graphical interface in TableCurve2D particularly useful. We recommend standardising the age-specific data to unit area before fitting. Once fitted, the predicted proportions can then be multiplied by the total flow to obtain the smoothed counts.

In Figure 13, we present eight model migration schedules fitted to the age-specific EU27 and LDC immigration flows for 2000, 2002, 2004 and 2006. The corresponding parameter values (along with 2008 values) are shown in Table 3. Finally, the observed data can be compared to the predicted data across five time points in Figure 14. The results show that the model migration schedules are useful for smoothing the age profiles of migration, whilst maintaining the overall pattern that would be expected.

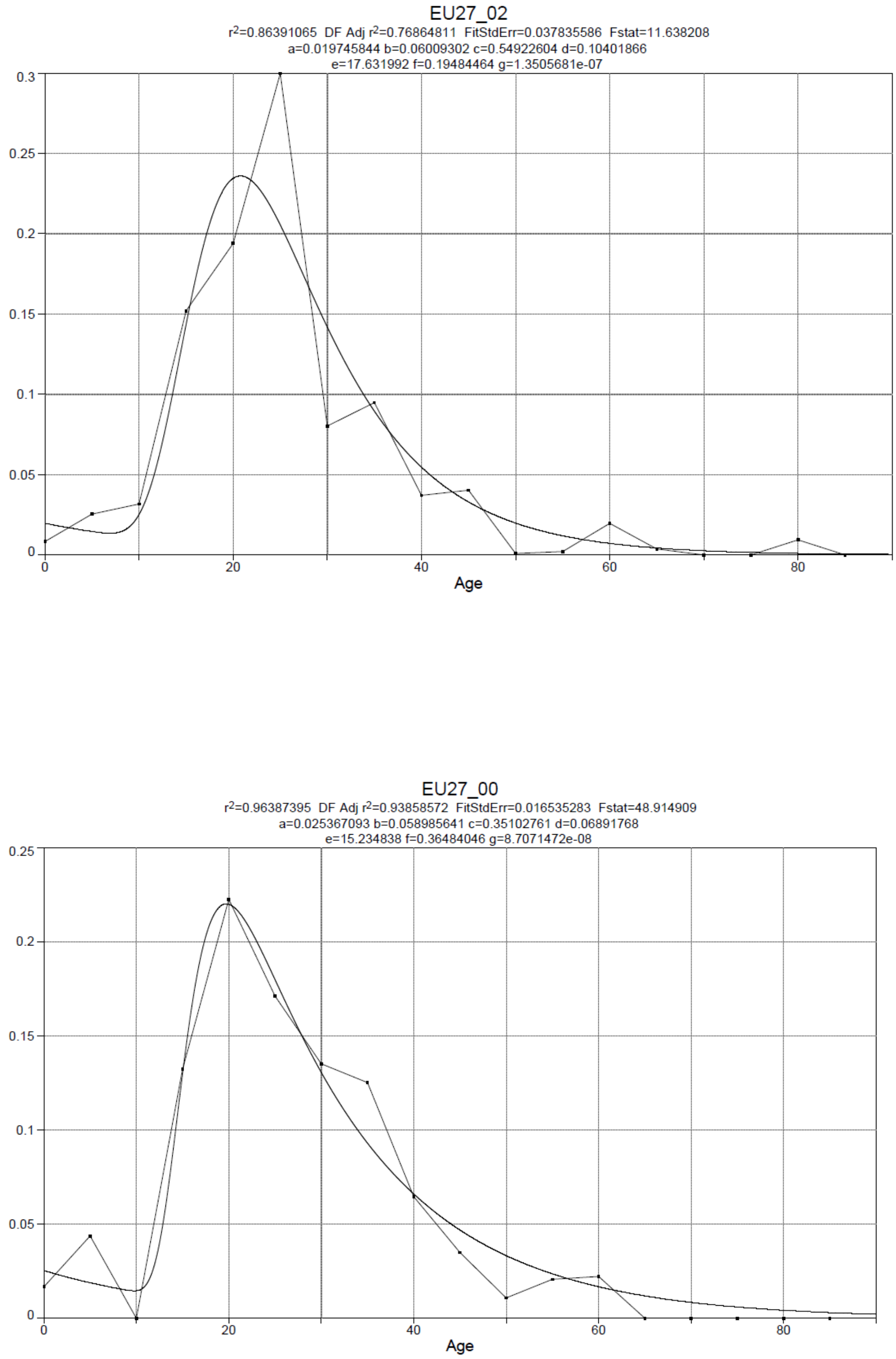
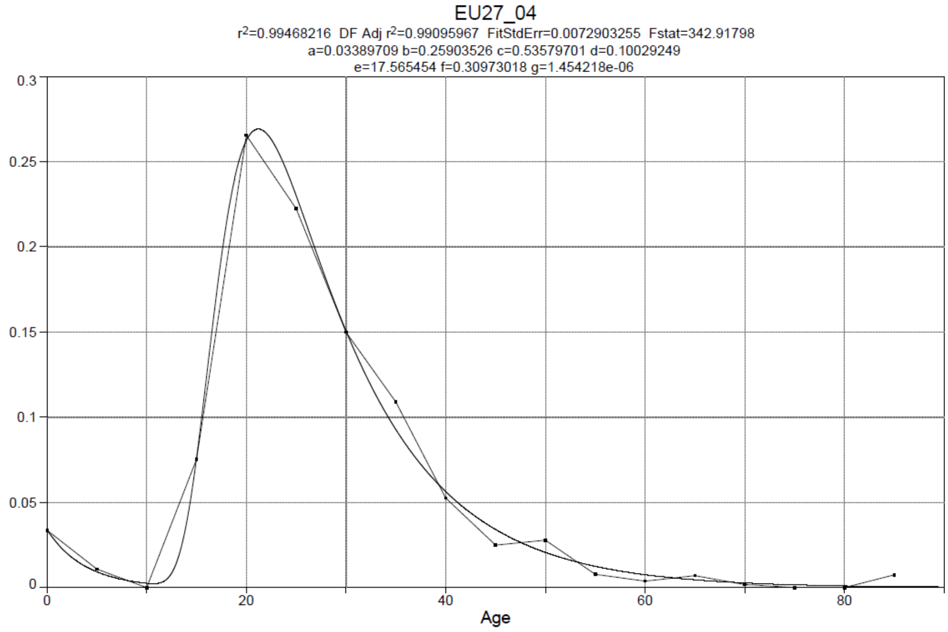


Figure 13a. Seven-parameter model migration schedules fitted to age compositions of immigration from EU27 countries, 2000 and 2002



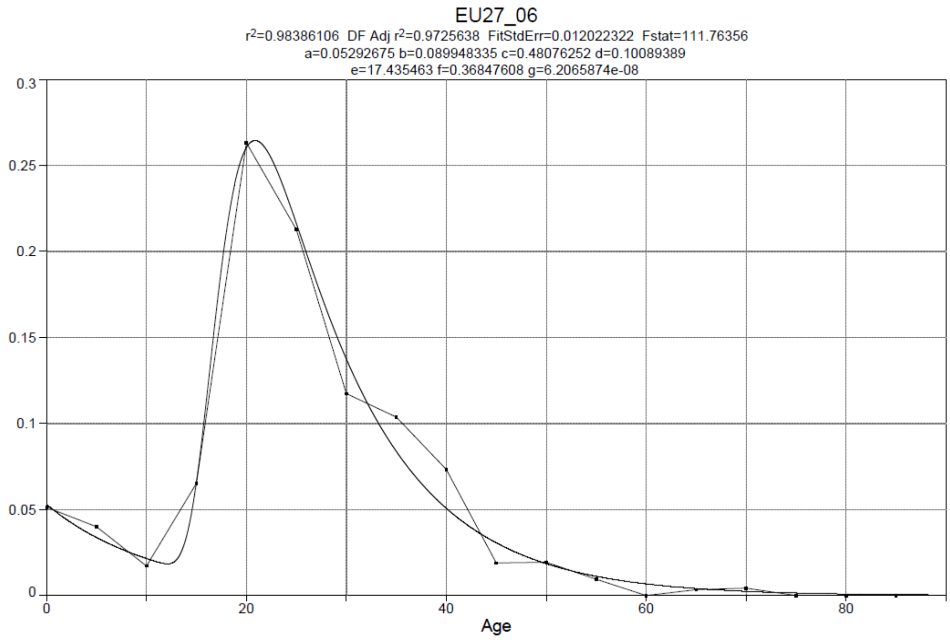
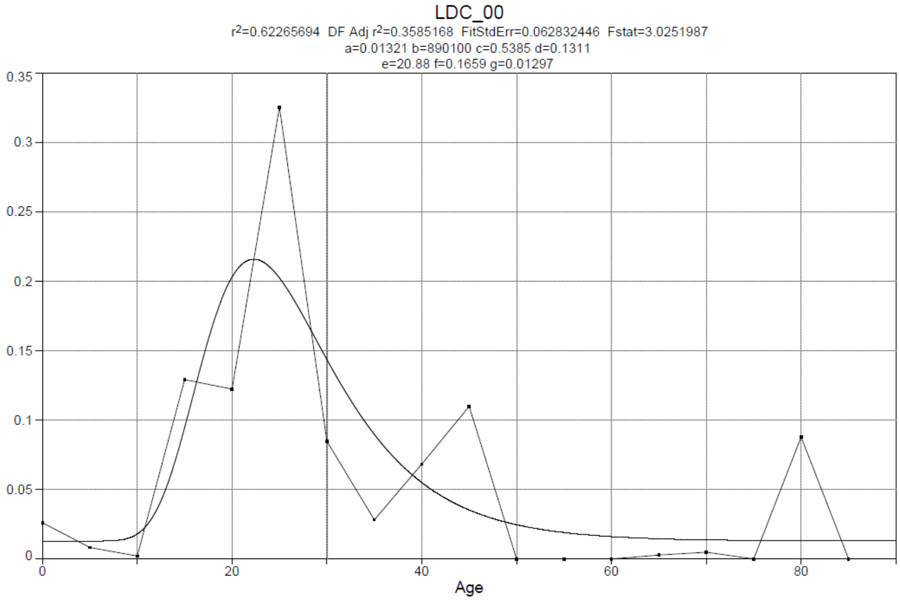


Figure 13b. Seven-parameter model migration schedules fitted to age compositions of immigration from EU27 countries, 2004 and 2006



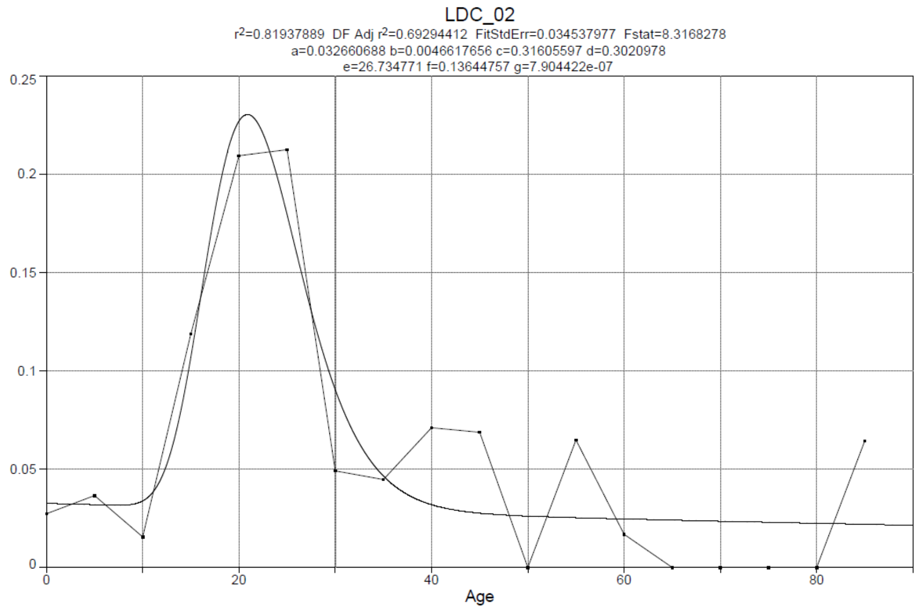
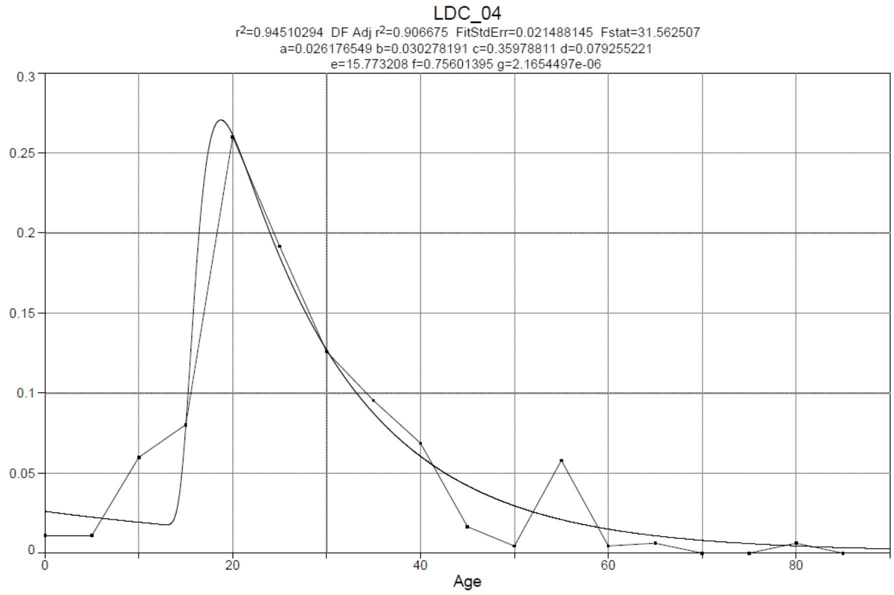


Figure 13c. Seven-parameter model migration schedules fitted to age compositions of immigration from LDC countries, 2000 and 2002



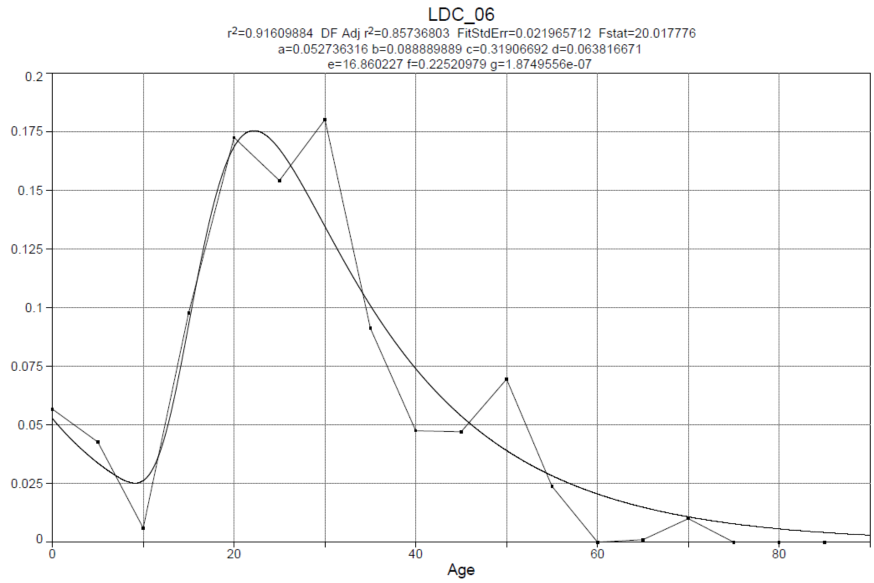


Figure 13d. Seven-parameter model migration schedules fitted to age compositions of immigration from LDC countries, 2004 and 2006

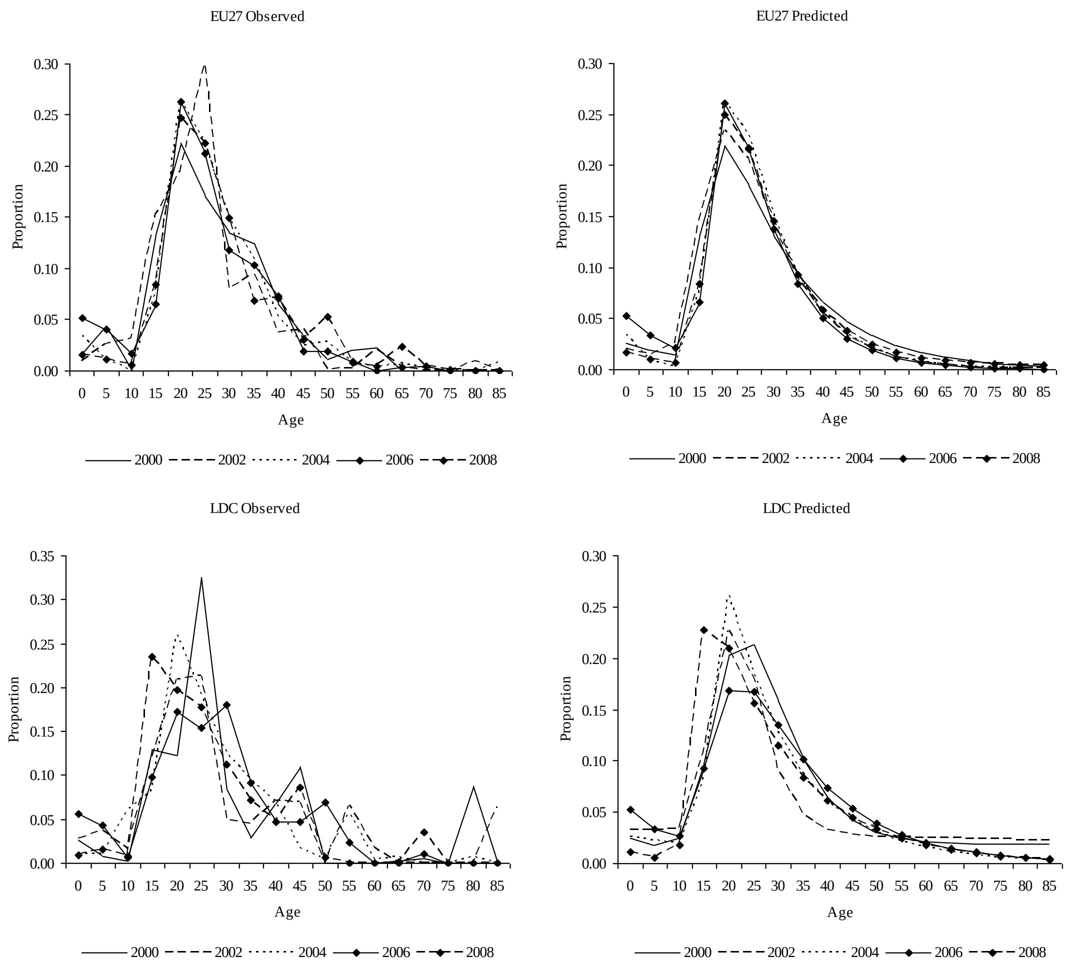


Figure 14. Comparison of observed and predicted age compositions of immigration from EU27 and LDC countries, 2002-2008

4.3 UNSATURATED LOG-LINEAR MODELS

Unsaturated log-linear models can be used to smooth the age and spatial structures in migration flow tables (Rogers et al. 2010, pp. 72-84). The model migration schedule approach described above can be considered as a “bottoms-up” approach that smoothes the age profile of each flow in a migration flow table. The log-linear model, on the other hand, can be viewed as a “top-down” approach in which higher-order marginal totals of, for example, an origin-by-age-by-sex table of migration flows are assumed to be more reliable (and regular) than lower-order marginal totals or cell values. Here, the data may be smoothed by removing, for example, the two-way and three-way interaction terms from the saturated model. Furthermore, model migration schedules may be combined with log-linear models to form hybrid models that may lead to further improvements in terms of both fit and parsimony (see Section 5).

Group	Parameter	2000	2002	2004	2006	2008
EU27	a_1	0.0254	0.0197	0.0339	0.0529	0.0127
	α_1	0.0590	0.0601	0.2590	0.0899	0.1370
	a_2	0.3510	0.5492	0.5358	0.4808	0.4921
	α_{12}	0.0689	0.1040	0.1003	0.1009	0.0967
	μ_2	15.2348	17.6320	17.5655	17.4355	17.3448
	λ_2	0.3648	0.1948	0.3097	0.3685	0.3088
	a_0	0.0000	0.0000	0.0000	0.0000	0.0036
	R^2	0.9639	0.8639	0.9947	0.9839	0.9797
LDC	a_1	0.0065	0.0327	0.0262	0.0527	0.0113
	α_1	0.9997	0.0047	0.0303	0.0889	0.1237
	a_2	0.5513	0.3161	0.3598	0.3191	0.3426
	α_{12}	0.1378	0.3021	0.0793	0.0638	0.0621
	μ_2	22.4451	26.7348	15.7732	16.8602	12.3232
	λ_2	0.1461	0.1364	0.7560	0.2252	0.5177
	a_0	0.0183	0.0000	0.0000	0.0000	0.0000
	R^2	0.6167	0.8194	0.9451	0.9161	0.9456

Table 3. Parameters and goodness-of-fit measures (R^2) for the seven-parameter model migration schedules fitted to age compositions of immigration from EU27 and LDC countries, 2002-2008

Consider the citizenship group data presented in Section 3.1. Each year, Eurostat requires a three-way table of immigration flows by citizenship group (C), age (A) and sex (S). A saturated log-linear model of this data for a single year is specified as

$$\log(n_{kxy}) = \lambda + \lambda_k^C + \lambda_x^A + \lambda_y^S + \lambda_{kx}^{CA} + \lambda_{ky}^{CS} + \lambda_{xy}^{AS} + \lambda_{kxy}^{CAS}, \tag{2}$$

where the subscripts k , x and s denote citizenship group, age group and sex, respectively. This model contains as many parameters as there are cell counts and, thus, predicts the data perfectly. What is important to note with this saturated model are the various structures contained within it. There are three main effects, three two-way interaction effects and one three-way interaction effect. This table of flows can be smoothed by removing various two-way and three-way interaction terms. For example, a main effects model, denoted C, A, S, is

$$\log(n_{kxy}) = \lambda + \lambda_k^C + \lambda_x^A + \lambda_y^S. \tag{3}$$

A model with a single two-way interaction term between citizenship group and age, denoted CA, S is specified as

$$\log\left(n_{kxy}\right) = \lambda + \lambda_k^C + \lambda_x^A + \lambda_y^S + \lambda_{kx}^{CA}, \tag{4}$$

and so forth.

The full set of unsaturated log-linear models starting with a main effects model are listed, along with likelihood ratio and Pearson Chi-Square measures of fit, in Table 4. Here, we see that the all two-way interaction model (i.e., CA, CS, AS) fits the IPS data the best, according to the likelihood ratio and Pearson chi-square statistics. However, this does not necessarily guarantee good results as demonstrated in Figure 15, where we see that the main effects (C, A, S) and two-way interaction model (CS, AS) models produce the most reasonable results. The models with the interaction between citizenship group and age are problematic because they contain zero values and irregularities, particularly for the smaller groups, such as the EFTA and LDC groups.

Model	Likelihood Ratio	Pearson Chi-Square	df
C, A, S	145,085	164,750	227
CA, S	51,999	46,420	125
CS, A	141,072	160,537	221
AS, C	134,574	144,248	210
CA, CS	47,986	42,854	119
CA, AS	41,488	38,176	108
CS, AS	130,560	139,681	204
CA, CS, AS	36,558	34,049	102

Table 4. Unsaturated log-linear model fits: Citizenship group (C) by age (A) by sex (S), 2009

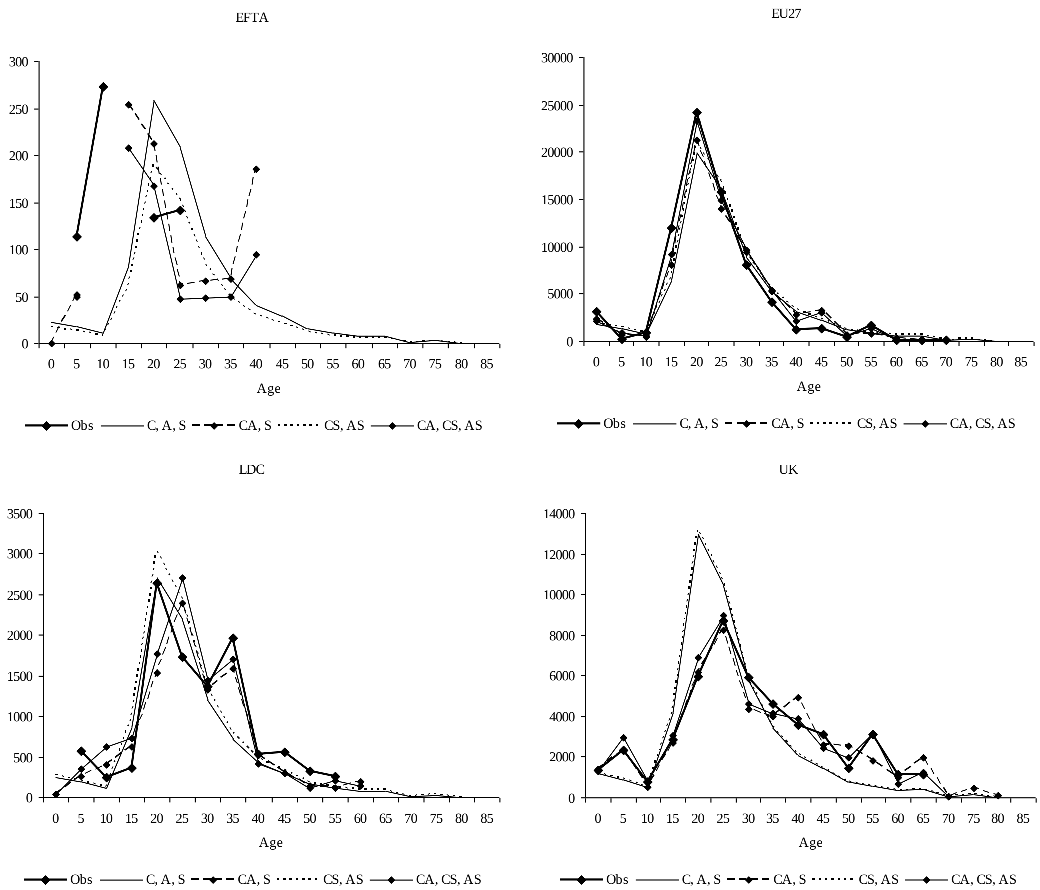


Figure 15. Comparison of observed and unsaturated log-linear predictions of immigration by citizenship group (C), age (A) and sex: Females, 2008

A reasonable model, considering the poor quality of the data, would be the (CS, AS) model. The results of applying this model to the IPS 2008 immigration by citizenship group, age and sex is presented in Figure 16 for females only. Here, we see that a single female age profile of migration is applied to all flows. The levels of the age profiles are set by the main effects and the two-way interaction between citizenship group and sex.

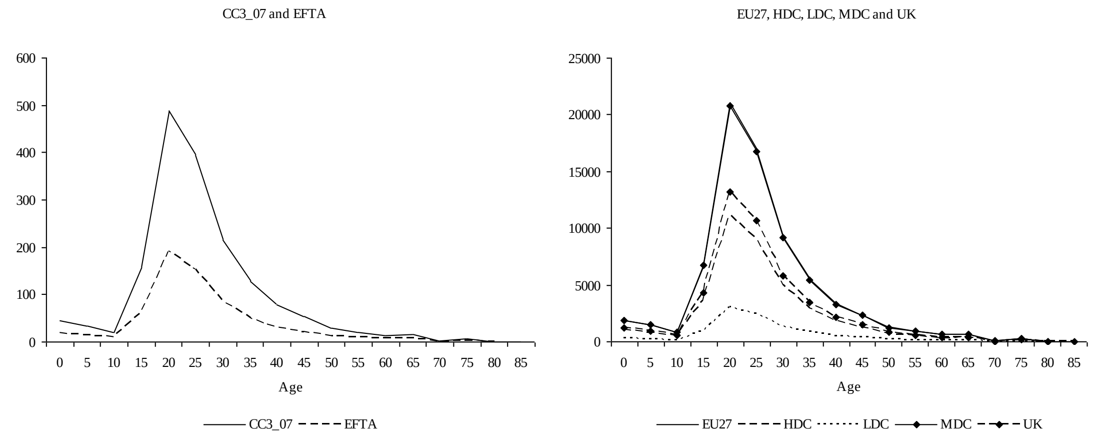


Figure 16. Unsaturated log-linear predictions of immigration by citizenship group (C), age (A) and sex: CA, AS model, females, 2008

Ideally, the interaction between citizenship group and age would be included to capture the likely different age profiles of, for example, returning UK nationals and entering LDC citizens. Unfortunately, the sample size of the IPS is too small for this. One way to overcome this would be to borrow strength over time (T) by including a time variable. This model is more complicated because it now has four dimensions. The saturated model for a citizenship group by age by sex by time table of immigration flows is specified as:

$$\begin{aligned}\log\left(n_{kxyt}\right) = \lambda + \lambda_k^C + \lambda_x^A + \lambda_y^S + \lambda_t^T + \lambda_{kx}^{CA} + \lambda_{ky}^{CS} + \lambda_{kt}^{CT} + \lambda_{xy}^{AS} + \lambda_{xt}^{AT} + \lambda_{yt}^{ST} \\ + \lambda_{kxy}^{CAS} + \lambda_{kxt}^{CAT} + \lambda_{kyt}^{CST} + \lambda_{xyt}^{AST} + \lambda_{kyt}^{CST} + \lambda_{xyt}^{AST} + \lambda_{kxyt}^{CAST}\end{aligned}, \qquad (3)$$

where the subscript t denotes year. For the purposes of this paper, we did not carry out this exercise as it is a straightforward extension of the three-way table illustration. Also, based on the pooled data analyses in Section 4.1, we know that this approach would not solve the problem with the two small citizenship groups of CC3 07 and EFTA. For these groups, no amount of smoothing would help. Instead we need to consider repairing or inferring methods.

5. REPAIRING METHODS FOR IMPROVING IPS DATA

We extend the unsaturated log-linear analysis in Section 4.3 to show how we can both smooth the reliable patterns and make assumptions to cover the unreliable patterns. Other repairing methods are not covered. These include borrowing patterns of migration from more reliable data, e.g., assuming EFTA age patterns are the same as for the EU27, and hierarchical disaggregation methods, which benchmarks the patterns considered reliable and assumes or predicts patterns for those that are not.

The multiplicative component model (Raymer and Rogers 2007; Raymer et al. 2011) is useful framework for repairing migration flows because, like the log-linear (statistical) model, it makes a distinction between an overall level, main effects, and interaction effects in contingency tables with parameters that can be used to guide the estimation process. This means that one can focus on modelling the underlying structures of migration flows via the multiplicative components. Also, the estimation process can be carried out in a systematic manner working from marginal effects to interaction effects. As described below, this model can also be extended to include other categorical variables, such as citizenship and sex. In fact, this modelling framework has been used in a variety of settings, for example, to project future age-specific migration patterns in Italy (Raymer et al. 2006), to combine migration data from multiple sources to study elderly and economic activity flows in England (Raymer et al. 2007 and Smith et al. 2010, respectively) and to construct missing origin-destination associations for migration between countries in Europe (Raymer et al. 2011).

For an illustration on how the multiplicative component model can be used to repair migration data, consider a simple two-way immigration table by citizenship group and age for 2009, which are presented in Table 5 for the observed IPS data. The multiplicative component model for this table is specified as:

$$n_{kx} = (T)(C_k)(A_x)(CA_{kx}), \tag{4}$$

where n_{kx} is an immigration flow of citizenship group k in age group x. There are four multiplicative components in total: an overall level, two main effects and one two-way interaction or association component. The multiplicative components are calculated with reference to the total level in the migration flow tables. The T component represents the total number of migrants in the system. The main effect components, C_k and A_x, represent proportions of all migration in each citizenship group and in each age group, respectively. The two-way interaction component represents the ratio of observed migration to expected migration (for the case of no interaction) and is calculated as $CA_{kx} = n_{kx} / [(T)(C_k)(A_x)]$. The CA_{kx} components represent the deviations from the overall age profile of migration, A_x. For estimation purposes, it is useful to know that they also represent ratios of the age compositions of citizenship groups to the overall age composition of migration, A_x.

The multiplicative components for the data presented in Table 5 are set out in Table 6. The overall level is presented in the bottom right corner (i.e., 528,094). The main effects for citizenship and age are presented in the bottom row and right column, respectively. Finally, the citizenship-age interaction components are presented within the margins of the table. For example, the observed 67,707 immigrants with MDC citizenship in age group 20-24 (see Table 5) can decomposed into the following four multiplicative components (see Table 6):

$$\begin{aligned} n_{6,20} &= (T)(C_6)(A_{20})(CA_{6,20}) \\ &= (528{,}094)(0.33562)(0.28664)(1.33275). \\ &= 67{,}707 \end{aligned}$$

The multiplicative components tell us that there were 528 thousand immigrants, of which 34 percent were MDC nationals and 29 percent were aged 20-24 years. Furthermore, the interaction term informs us that there were 33 percent more immigrants in this citizenship and age group than expected.

Age	CC3_07	EFTA	EU27	HDC	LDC	MDC	UK	Total
				Citizenship Group				
0	0	0	5,164	3,428	79	2,044	3,021	13,737
5	0	113	1,519	1,436	573	1,943	5,001	10,585
10	0	0	886	1,342	885	1,963	1,115	6,192
15	287	550	17,545	6,787	1,342	16,588	5,812	48,911
20	457	460	46,024	20,032	3,320	67,707	13,370	151,370
25	924	134	30,216	19,917	5,182	47,729	17,868	121,970
30	620	142	20,309	11,045	2,875	22,552	9,490	67,033
35	0	150	11,516	5,891	3,441	10,223	8,656	39,877
40	0	401	5,949	2,810	1,115	3,230	10,700	24,206
45	0	0	7,026	1,842	658	1,709	5,672	16,907
50	0	0	1,556	1,114	319	678	5,532	9,199
55	0	0	1,636	484	256	222	3,949	6,547
60	0	0	793	713	430	401	2,108	4,445
65	0	0	328	0	0	0	4,303	4,631
70	0	0	141	73	0	247	61	522
75	0	0	0	707	0	0	1,054	1,761
80	0	0	0	0	0	0	202	202
85	0	0	0	0	0	0	0	0
Total	2,288	1,950	150,609	77,622	20,476	177,237	97,913	528,094

Table 5. Observed immigration by age and citizenship group, 2009
Source: International Passenger Survey

Age	CC3_07	EFTA	EU27	Citizenship HDC	LDC	MDC	UK	Total
0	0.00000	0.00000	1.31813	1.69781	0.14890	0.44338	1.18627	0.02601
5	0.00000	2.89819	0.50325	0.92319	1.39532	0.54690	2.54807	0.02004
10	0.00000	0.00000	0.50170	1.47506	3.68703	0.94481	0.97113	0.01172
15	1.35408	3.04531	1.25782	0.94402	0.70787	1.01053	0.64085	0.09262
20	0.69721	0.82296	1.06612	0.90035	0.56568	1.33275	0.47640	0.28664
25	1.74762	0.29709	0.86865	1.11098	1.09578	1.16597	0.79013	0.23096
30	2.13582	0.57450	1.06231	1.12100	1.10608	1.00244	0.76356	0.12693
35	0.00000	1.01765	1.01263	1.00509	2.22564	0.76385	1.17071	0.07551
40	0.00000	4.48534	0.86183	0.78984	1.18853	0.39760	2.38413	0.04584
45	0.00000	0.00000	1.45712	0.74132	1.00364	0.30124	1.80931	0.03202
50	0.00000	0.00000	0.59307	0.82350	0.89519	0.21970	3.24352	0.01742
55	0.00000	0.00000	0.87629	0.50298	1.00766	0.10105	3.25323	0.01240
60	0.00000	0.00000	0.62538	1.09081	2.49379	0.26904	2.55829	0.00842
65	0.00000	0.00000	0.24866	0.00000	0.00000	0.00000	5.01103	0.00877
70	0.00000	0.00000	0.94597	0.95773	0.00000	1.40777	0.63090	0.00099
75	0.00000	0.00000	0.00001	2.73058	0.00000	0.00000	3.22879	0.00334
80	0.00000	0.00000	0.00004	0.00000	0.00000	0.00000	5.39345	0.00038
85	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Total	0.00433	0.00369	0.28519	0.14699	0.03877	0.33562	0.18541	528,094

Table 6. Observed multiplicative components of immigration by age and citizenship group, 2009

In terms of repairing the data, let's assume that the overall level and main effect components, shown in Figure 17, are reliable and that the CA_{kx} interaction terms are in need of repair. In examining the age patterns of the seven citizenship groups, we find that the age patterns of the five larger flows could benefit from being smoothed with model migration schedules. The patterns for the two smaller flows (CC3 07 and EFTA) need to be imposed.

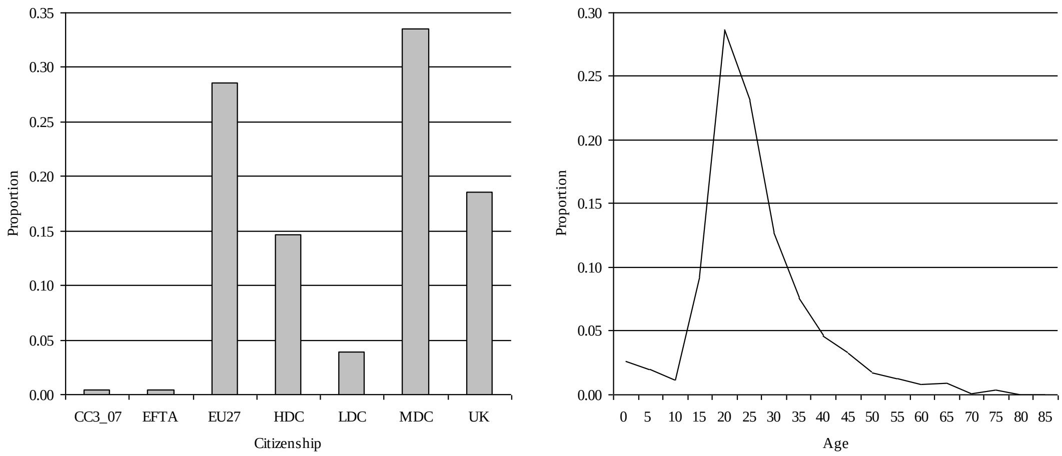


Figure 17. The proportions of immigration by citizenship group and age, 2009

To repair the citizenship group by age interactions, we first fit model migration schedules to the five reliable age compositions (standardised to unit area) of reported migration to smooth out minor irregularities. These schedules are presented in Figure 18. We then divided these age compositions by a model schedule fit to the overall age composition of migration (i.e., the A_x component) to obtain estimates of the CA_{kx} components for these five flows. Note, the A_x component was smoothed primarily to remove the minor irregularities in the oldest age groups. Finally, we set the ratios for the two small citizenship groups to one. By setting these ratios to one, we are assuming the age profiles of these flows correspond to the age profile in the age main effect (i.e., the average age profile observed). (Alternatively, we could have set them equal to one of the other five larger groups, e.g., EU27). The predicted ratios are presented in Table 7, along with the main effect and overall level components.

Once the multiplicative components are obtained, we can then estimate an initial (unconstrained) set of immigration flows by citizenship and age. These flows are set out in Table 8. To constrain the estimates to the original marginal totals, one can simply rescale these numbers to the marginal totals in Table 5 by using iterative proportional fitting or a log-linear with offset model (described in the next section). Our final repaired immigration data results, with marginal totals matching those in Table 5, are presented in Table 9.

Age	Citizenship CC3_07	EFTA	EU27	HDC	LDC	MDC	UK	Total
0	1.00000	1.00000	1.74762	2.19801	1.45724	0.68936	2.03521	0.01942
5	1.00000	1.00000	0.61327	1.11674	1.48917	0.56657	1.80085	0.01847
10	1.00000	1.00000	0.21391	0.77533	1.51336	0.46283	1.58604	0.01767
15	1.00000	1.00000	1.26673	0.92437	0.43905	0.70924	0.63853	0.09163
20	1.00000	1.00000	1.20186	1.00669	0.67962	1.40870	0.61370	0.25246
25	1.00000	1.00000	0.97456	1.16372	1.10396	1.07422	0.77095	0.21151
30	1.00000	1.00000	0.96457	1.10745	1.15407	1.00876	1.00223	0.13155
35	1.00000	1.00000	0.99318	0.91986	1.24327	0.99081	1.28377	0.07809
40	1.00000	1.00000	0.99328	0.72932	0.88782	0.94811	1.64217	0.04770
45	1.00000	1.00000	0.93501	0.60133	0.92462	0.85700	1.82244	0.03096
50	1.00000	1.00000	0.81476	0.55002	1.08188	0.72049	1.95178	0.02170
55	1.00000	1.00000	0.65493	0.56037	1.29570	0.56220	1.93276	0.01650
60	1.00000	1.00000	0.49008	0.60643	1.50698	0.41161	1.78654	0.01347
65	1.00000	1.00000	0.34705	0.67715	1.65865	0.32658	1.63054	0.01162
70	1.00000	1.00000	0.23659	0.72685	1.75744	0.19646	1.32820	0.01041
75	1.00000	1.00000	0.15742	0.78451	1.81745	0.11535	1.10229	0.00956
80	1.00000	1.00000	0.10321	0.83436	1.85204	0.06674	0.90479	0.00891
85	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.00838
Total	0.00433	0.00369	0.28519	0.14699	0.03877	0.33562	0.18541	528,094

Table 7. Estimated multiplicative components of immigration by age and citizenship group, 2009

Age	CC3_07	EFTA	EU27	Citizenship HDC	LDC	MDC	UK	Total
0	44	38	5,111	3,313	579	2,373	3,870	15,328
5	42	36	1,706	1,601	563	1,854	3,256	9,058
10	40	34	569	1,063	547	1,449	2,743	6,447
15	210	179	17,481	6,575	824	11,518	5,729	42,515
20	578	492	45,697	19,727	3,513	63,032	15,170	148,209
25	484	412	31,045	19,106	4,781	40,270	15,966	112,065
30	301	256	19,110	11,308	3,109	23,519	12,909	70,512
35	179	152	11,680	5,575	1,988	13,713	9,815	43,102
40	109	93	7,136	2,700	867	8,015	7,670	26,591
45	71	60	4,360	1,445	586	4,702	5,524	16,748
50	50	42	2,663	927	481	2,772	4,148	11,082
55	38	32	1,627	718	438	1,644	3,122	7,618
60	31	26	994	634	416	983	2,356	5,440
65	27	23	607	611	395	673	1,855	4,190
70	24	20	371	588	375	363	1,354	3,095
75	22	19	227	582	356	196	1,032	2,433
80	20	17	139	577	338	105	789	1,986
85	19	16	1,262	650	172	1,485	820	4,424
Total	2,288	1,950	151,786	77,700	20,326	178,664	98,129	530,844

Table 8. Initial (unconstrained) repaired immigration flows by age and citizenship group, 2009

Age	CC3_07	EFTA	EU27	Citizenship HDC	LDC	MDC	UK	Total
0	40	34	4,485	2,964	543	2,084	3,587	13,737
5	49	42	1,938	1,854	683	2,109	3,908	10,583
10	39	33	529	1,007	543	1,348	2,693	6,192
15	245	209	19,864	7,616	999	13,102	6,875	48,910
20	601	512	46,183	20,325	3,790	63,767	16,192	151,370
25	535	455	33,296	20,890	5,474	43,234	18,086	121,970
30	290	247	17,864	10,777	3,102	22,008	12,745	67,033
35	167	142	10,605	5,161	1,927	12,463	9,412	39,877
40	100	85	6,363	2,455	825	7,154	7,223	24,205
45	72	61	4,301	1,454	617	4,644	5,757	16,906
50	41	35	2,155	764	415	2,244	3,544	9,198
55	32	28	1,359	611	390	1,374	2,753	6,547
60	25	21	787	512	351	779	1,970	4,445
65	29	25	649	665	450	719	2,093	4,630
70	4	3	60	97	65	59	233	521
75	16	13	158	414	265	136	760	1,762
80	2	2	14	57	35	10	81	201
85	0	0	0	0	0	0	0	0
Total	2,287	1,947	150,610	77,623	20,474	177,234	97,912	528,087

Table 9. Repaired immigration flows by age and citizenship group, 2009

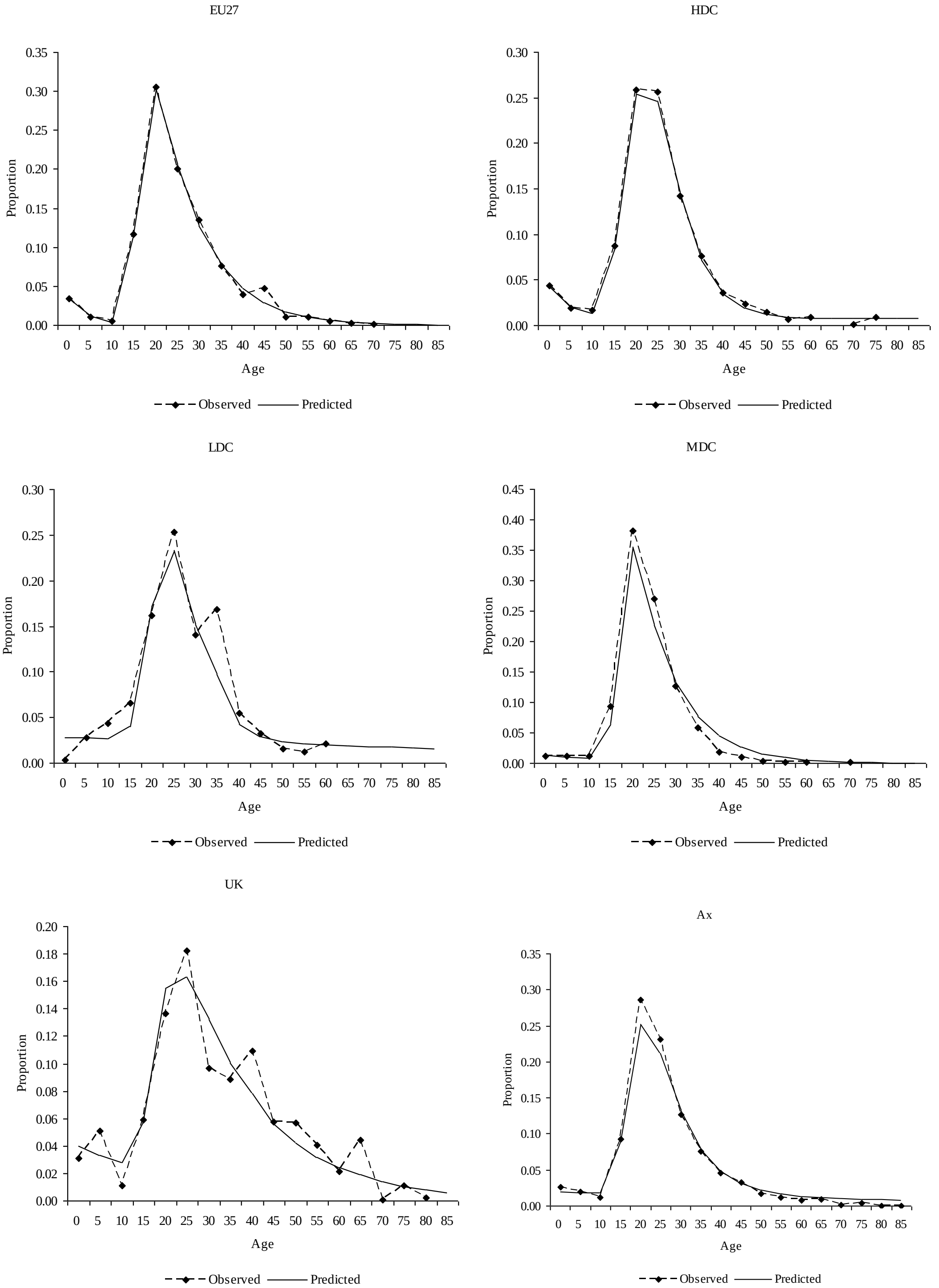


Figure 18. Model schedule fits to age compositions of immigration by citizenship group and to the overall age profile of migration (A_x), 2009

6. INFERRING METHODS FOR IMPROVING IPS DATA

In this section, we focus on inferring methods for improving the IPS data. Three approaches are introduced. The first combines higher education data with the IPS data to estimate the origin, age and sex patterns of immigration. The second approach applies regression methods to estimate the origins of immigrants based on IPS data, pooled over ten years, and covariate information. Finally, the third approach combines migration data collected by sending and receiving countries throughout Europe to estimate origin-destination-specific flows.

6.1 INCORPORATING AUXILIARY INFORMATION

To illustrate the incorporation of auxiliary information, we combine IPS data on migration flows by broad age group, country of previous residence and sex (IMM5PVR) for 2000-2007 with corresponding counts of foreign students in Higher Education institutions, maintained by the Higher Education Statistics Agency (HESA). Due to confidentiality agreements with HESA, the results from this work are not presented in detail.

The number of migrants aged 20-24 in 2007 reported by the IPS and HESA data sources were compared for the top 20 student origins. We found that there were some large differences in the totals, most notably from Poland, whose flows were typically for reasons other than education. For flows from smaller countries, HESA figures are generally larger than estimates from IPS. This is believed to be associated with the better coverage of the HESA data, collected from enrolled students at higher education institutes. For other countries with even smaller flows, there are many situations where the HESA data report flows of foreign students while the IPS reports zeros.

The comprehensive origin structure found in the HESA data may be beneficial in estimating detailed migration flow counts from country-specific origins, where flows are dominated by student migrants. This can be undertaken in the log-linear model framework, using the origin structure from the HESA data as auxiliary information, via an offset term. For example, consider a log-linear model that includes age, sex and the age-sex interaction covariates:

$$\log\left(n_{ixy}\right) = \lambda + \lambda_x^A + \lambda_y^S + \lambda_{xy}^{AS} + \log\left(y_{ixy}\right), \tag{4}$$

where the observed IPS data for each origin-age-sex is denoted as n_{ixy}, and y_{ixy} denotes the corresponding HESA data. The offset term imposes the origin structure of the HESA data on the predicted values, which are constrained to the IPS overall level and age-sex distributions.

The fitted age schedules from the log-linear model reflected a more classical age schedule pattern in comparison to the raw IPS data. They also tended to follow the broader patterns discussed in Section 3, including wider labour force peaks for males. For flows from countries that have large known student populations in the United Kingdom, such as Chinese males and females, Taiwanese females and Greek males, the fitted values extended the peak of age schedules well above that recorded by the IPS. In cases where the flows were not strongly related to educational factors, such as Indian females, the fitted values shrinked the peak of the age schedules below that recorded from the IPS. This resulted from the inclusion of the offset term based on HESA to dictate the origin structure of all migration flows, which may or may not be related to education.

The tendency for under-estimating migration flows from countries with immigrants moving for non-educational reasons could be alleviated by augmenting the HESA data with counts of non-student flows from other sources, such as the 2001 and 2011 censuses or new National Insurance Number registrations of persons born abroad. Moreover, migrants by stated reason of entry (e.g., for study, family reunion or work) could be modelled separately as Boden and Rees (2010) proposed for subnational estimation of immigration.

6.2 MODEL-BASED ESTIMATION

A model-based approach for estimating the international migration flows to the United Kingdom may also be used to estimate migration flows. This approach has been used, for example, by Abel (2010) to estimate the missing flows within EU-15 countries and by Raymer et al. (2011) to estimating missing flows in the MIgration MOdelling for Statistical Analyses (MIMOSA) project (see also de Beer et al., 2010).

For illustration of the model-based approach, we use data on total immigration flows by country of previous residence (IMM5PRV), aggregated over time from 2000 to 2009. Further aggregation by groups of countries is undesirable as it reduces the number of observations substantially. It is assumed that zero flows (for 45 countries) are not observed due to the small sample of the IPS; they are treated as missing data and are excluded from the estimation. The dependent variable is a logarithm of immigration flows, y_i. The equation can be written as:

$$\log y_i = \alpha_0 + \alpha_1 \log P_i + \alpha_2 EL_i + \alpha_3 VR_i + \alpha_4 EU_i + \alpha_5 \log D_i + \alpha_6 BO_i + \alpha_7 BC_i + \varepsilon_i, \quad (5)$$

where ε_i is normally distributed with variance σ^2 and the covariates used for estimation are:

- P_i – population size of the sending country (logged, source: Population Reference Bureau's World Population Sheet 2010).
- EL_i – a dummy for English speaking countries, according to Drinkwater (2006), countries: Australia, Canada, Ireland, the United States, New Zealand and South Africa.
- VR_i – a dummy for visa requirement (countries list according to UK Border Agency, http://www.ukba.homeoffice.gov.uk/policyandlaw/immigrationlaw/immigrationrules/appendix1/, accessed in March 2011).
- EU_i – a dummy for EU-27 or EFTA country.
- BO_i – a dummy for British overseas territory (countries according to UK Boarder Agency, http://www.ukba.homeoffice.gov.uk/britishcitizenship/othernationality/britishoverseasterritories/, accessed in March 2011).

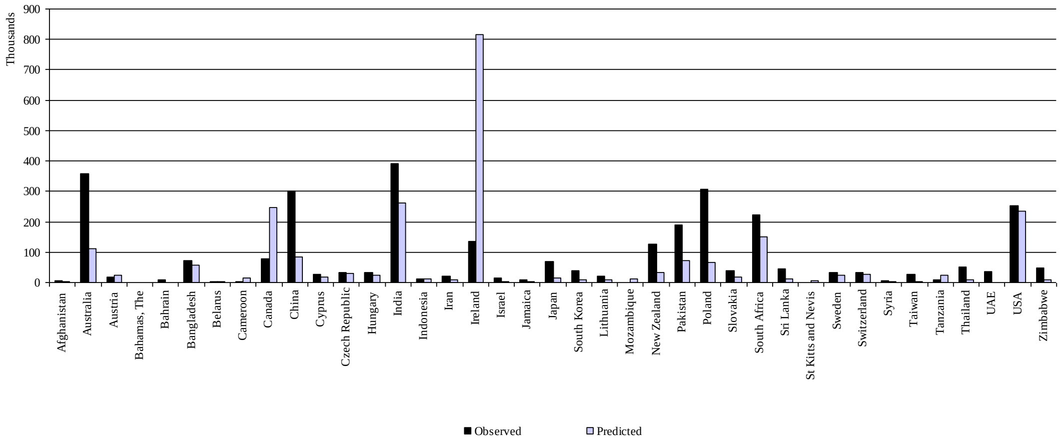
- BC_i – a dummy for British Commonwealth present and former members (55 countries, according to the Commonwealth Secretariat, http://www.thecommonwealth.org/Internal/142227/members/, accessed in March 2011). A second version of the model (described below) assumes only present members of the Commonwealth (without Ireland and Zimbabwe).
- D_i – a weighted distance between the UK and the countries of origin, obtained from Mayer and Zignago (2006).

The model is estimated using the Ordinary Least Squares (OLS) method available in all statistical software, as well as in the spreadsheet programme *Excel*. All variables apart from the constant are significant with p-values lower than 0.05. The estimation results of the model with the Commonwealth dummy including present and former members are presented in Table 10. The signs of the coefficients are consistent with expectations. The adjusted R-square is 0.58, which means that the model explains around 58% of the variability in the reported migration flows. Hence, the model fits the data reasonably well. Hypothesis about homoscedasticity of errors is not rejected using White test and normality of errors is confirmed by Jarque-Bera test. Note, that this model is for illustration purposes only. If used in practice, extensions should include more economic, demographic and geographic covariates.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.21	1.82	-0.12	0.9070
Population	0.73	0.06	11.68	0.0000
English speaking	1.62	0.61	2.66	0.0087
Visa requirement	-0.57	0.27	-2.08	0.0388
EU27	1.53	0.47	3.24	0.0015
Distance	-0.41	0.19	-2.17	0.0316
British overseas	2.84	0.64	4.45	0.0000
Commonwealth	1.63	0.27	6.15	0.0000
R-square = 0.58				

Table 10. Results of OLS model estimation

A. Ireland included in list of Commonwealth countries



B. Ireland not included in list of Commonwealth countries

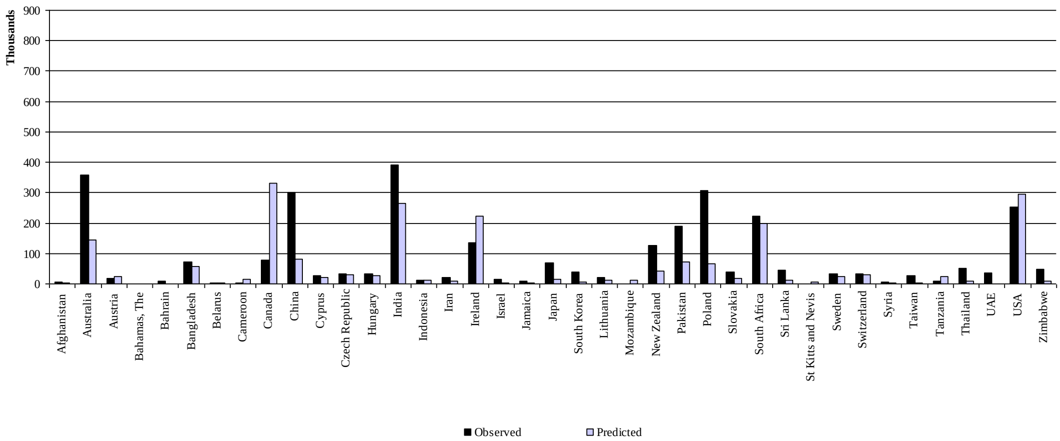


Figure 19. Observed and predicted flows based on OLS regression: A comparison of the results based on different assumptions regarding Ireland, total flows 2000-2009

The predicted values are used as estimates for the immigration flows. Model-based predictions are also made for countries with zero flows measured. A comparison of flows for selected countries for two versions of the model is presented in Figure 19. In the upper figure, the Commonwealth indicator variable includes both former and current members, whilst in the bottom figure, only current members are included.

It can be noticed that for some countries the predicted values are smaller than the observed ones. For example, the observed flows from India are nearly 400k while model predicts 280k immigrants for the period 2000-2009. Some of the extreme cases are Poland (300k versus 60k predicted), Australia (350k vs. 100k predicted) or China

(300k vs. 100k predicted). This may result from lack of more explanatory variables, e.g. economic (GDP, GNI, unemployment) or demographic (life expectancy, age dependence ratio). For Ireland, for which the measurement is available only for years 2008 and 2009, the imputation of the mean of these data for years 2002-2007 is applied. Thus the reported flow used in estimation is 130k instead of original 26k, resulting in a predicted value of 800k. This results mainly from the very short distance between Ireland and the UK and the fact that it used to be a part of the Commonwealth. If it is excluded from the Commonwealth (the dummy is equal to zero), it only slightly affects the model parameters and flows from the other countries. However, the flow from Ireland is reduced to 220k, which seems to be a more plausible number.

The total count of migrants estimated by the IPS survey is 4.8M. The model predicts 3.8M migrants (including countries for which there is no reported flow). A version of the model with Ireland excluded from the Commonwealth results in 3.5M inflows predicted. Out of 220 predicted flows, 119 are larger than the reported. That includes 45 countries, where no flows are observed. The total flows from these countries are almost 26k people. The origin structure predicted by the model can be applied to the total IPS number of migrants. This results in a 26% increase in all estimates together (i.e., $4.8/3.8 \cong 1.26$).

A similar model to the one described above can be fitted separately to males and females. However, this is problematic due to the relatively small IPS samples. In many situations, e.g., Slovenia or Bosnia and Herzegovina, flows for only one sex are observed. One solution would be to build a model for total flows and then redistribute them to both sexes, possibly using some smoothing algorithms or borrowing structure from the other sources. Another option is to obtain a breakdown by origin from the separate models for both sexes and then use aggregation. However, the latter approach is questionable due to the suspicious IPS estimates of the flows data for males and females (see Section 3).

Summarising, the model-based estimation can serve as a tool for obtaining the origin structure of the immigration flows. The approach presented in this section can be extended in several ways. One of them is using a model for males and females separately. Secondly, additional covariates, such as economic or demographic, can be included in the explanatory part. Third option is to use a panel approach for all years and treat the zero counts as missing data, which can be later predicted. However, this would require using more advanced estimation techniques.

6.3 MIMOSA / IMEM / ABEL APPROACHES

Recently, there have been three projects on estimating international migration flows amongst countries in Europe. The first is the MIgration MOdelling for Statistical Analyses (MIMOSA) project[3], which was funded by Eurostat to estimate international migration stocks and flows in Europe. The methodological work on estimating flows is described in de Beer et al. (2010) and Raymer et al. (2011). The second is the currently on-going Integrated Modelling of European Migration (IMEM) project, funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)[4]. An introduction to this project can be found in Raymer et al. (2010). Both the MIMOSA and IMEM projects rely on the data provided by sending and receiving countries in Europe. The third project represents Guy Abel's PhD work on 'International Migration Flow Table Estimation' (see Abel 2010).

The methodology adopted by the MIMOSA team represents a two-stage hierarchical procedure. The first stage harmonises the available immigration and emigration data by using a simple optimisation procedure (Poulain 1999) benchmarked to Sweden's migration flow data, which are assumed to be measured more or less without error (see also de Beer et al. 2009). The second stage estimates the missing marginal data and associations between countries by using the available flows and covariate information. Both stages are set within a multiplicative framework for analysing migration flows. No measures of uncertainty are provided and the approach is sensitive to the model assumptions and estimation procedure.

The IMEM project utilises a Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows.

[3] http://www.nidi.knaw.nl/Pages/NID/24/928.bGFuZz1VSw.html
[4] http://www.norface.org/migration12.html

The focus is on estimating recent international migration flows between countries in the European Union, using data primarily collected by Eurostat and other national and international institutions, as well as qualitative information from experts. The methodology is integrated and capable of providing a synthetic data base with measures of uncertainty for international migration flows and other model parameters.

The results of the MIMOSA and IMEM projects, as well as those produced in Abel (2010), provide a base for countries to compare and improve their statistics on migration as required in the 2007 regulation on migration statistics passed by the European Parliament (see below). The methodologies are based on the idea of combining data obtained from multiple countries. ONS could benefit from this approach, at the very least, by comparing their estimated figures of, say, immigration from Germany with Germany’s emigration figures. However, this will only help, if the user knows that Germany applies a relatively loose definition of migration and therefore its figures are higher than those using, say, a six month (e.g., Norway) or twelve month (e.g., Sweden) definition.

7. SUMMARY AND RECOMMENDATIONS

In 2007, the European Parliament passed a regulation to govern the supply of national statistics to the EU. Countries are now required to provide harmonised migration flow statistics to Eurostat in accordance to Regulation 862/2007.[5] Recognising the many obstacles with existing data, Article 9 of the Regulation states that 'As part of the statistics process, scientifically based and well documented statistical estimation methods may be used.' The methods introduced in this paper should help the Office for National Statistics satisfy the requirements set out in Article 3 of the 2007 Regulation.

We have illustrated various methods that can be used to improve or estimate multidimensional tables of IPS-based immigration flows. The results represent synthetic data benchmarked to IPS marginal totals that are deemed reliable. Furthermore, the methods can be readily extended to estimate emigration flows and

5 http://europa.eu/legislation_summaries/justice_freedom_security/free_movement_of_persons_asylum_immigration/l14508_en.htm

other multidimensional tables. The multiplicative component model framework (Sections 5 and 6) is particularly useful for combining reliable structures with smoothed, repaired or inferred structures. We advocate applying this approach.

While the illustrations presented in this paper are by no means perfect, we believe they provide a substantial and significant improvement over the patterns exhibited in the observed flows, which contain irregularities and missing data due to sample size. Our methodology is based on the idea of smoothing, repairing and combining data. Further investigation needs to be made on the model designs corresponding to each of Eurostat’s mandatory tables, including those for emigration flows.

Our recommendations for improving the UK’s immigration and emigration data to meet Eurostat’s requirements are as follows. First, for each required table, the reliable and unreliable structures should be identified for its particular theoretical multiplicative component model (i.e., the model that captures most of the patterns). In most cases, a two-way interaction model should suffice. However, there may be cases where three-way interactions are required. Second, where necessary, the reliable structures should be smoothed to remove unexpected irregularities due to the relatively small sample size of the IPS. Third, for the unreliable structures, there should be analyses undertaken to either repair these data or infer them based on auxiliary or covariate information (or both). Here, experts may be needed to assess the reasonableness of the estimated components, e.g., the proportion of immigrants by country of previous residence, and to help design the model. Finally, the (smoothed) reliable and estimated structures should be combined by using iterative proportional fitting or log-linear with offset models.

REFERENCES

Abel, G.J. (2010) Estimation of international migration flow tables in Europe. *Journal of the Royal Statistical Society Series A* (*Statistics in Society*) 173(4), 797-825.

Boden, P. and Rees, P. (2010) Using administrative data to improve the estimation of immigration to local areas in England. *Journal of Royal Statistical Society*173 (4): 707-731

de Beer, J., Raymer, J., van der Erf, R. and van Wissen, L. (2010) Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *European Journal of Population* 26(4):459-481.

Drinkwater, S., Eade, J. and Garapich, M. (2006) Poles Apart? EU Enlargement and the Labour Market Outcomes of Immigrants in the UK. IZA Discussion Paper No. 2410, Institute for the Study of Labor, Bonn, Germany.

Mayer, T. and S, Zignago. (2006) Notes on CEPIIs distances measures. Centre d'Etudes Prospectives d'Informations Internationales (CEPII), Paris.

Poulain, M. (1999), International migration within Europe: towards more complete and reliable data?. Working Paper 12, joint ECE-Eurostat Work Session on Migration Statistics, Geneva, Switzerland.

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

Raymer, .J and Bijak, J. (2009) Report of the technical consultancy in the UNITED KINGDOM on 22 May 2009. MIMOSA Deliverable 10.1A, Modelling of Statistical Data on Migration and Migrant Populations, Eurostat Project 2006/S 100-10667/EN LOT 2, Eurostat, Luxembourg.

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

Raymer, J., de Beer, J. and van der Erf, R. (2011) Putting the pieces of the puzzle together: Age and sex-specific estimates of migration amongst countries in the EU / EFTA, 2002-2007. *European Journal of Population* 27:185-215.

Raymer, J., Forster, J.J., Smith, P.W.F., Bijak, J., Wiśniowski, A. and Abel, G.J. (2010) The IMEM model for estimating international migration flows in the European Union. Working paper 14, Joint UNECE/Eurostat Work Session on Migration Statistics, Geneva.

Raymer, J. and Rogers, A. (2007) Using age and spatial flow structures in the indirect estimation of migration streams. *Demography* 44(2):199-223.

Rogers, A. and Castro, L.J. (1981) Model Migration Schedules. RR-81-30, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Rogers, A. and Little, J.S. (1994) Parameterizing age patterns of demographic rates with the multiexponential model schedule. *Mathematical Population Studies* 4(3):175-194.

Rogers, A., Little, J. and Raymer, J. (2010) *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. Springer: Dordrecht.

Smith, P.W.F., Raymer, J. and Giulietti, C. (2010) Combining available migration data in England to study economic activity flows over time. *Journal of the Royal Statistical Society Series A* (*Statistics in Society*) 173(4):733-753.

APPENDIX: LIST OF COUNTRIES ACCORDING TO COUNTRY GROUP

CC3_07	Croatia, Macedonia, Turkey
EFTA	Iceland , Liechtenstein, Norway, Switzerland
EU27	Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Czechoslovakia (Ex)
HDC	Andorra, United Arab Emirates, Antigua and Barbuda, Anguilla, Albania, Antilles / Curacao, Argentina, Australia, Aruba, Bosnia Herzegovina, Barbados, Bahrain, St Barthelemy, Bermuda, Brunei, Brazil, Bahamas, Belarus, Canada, Chile, Costa Rica, Cuba, Ecuador, Falkland Islands / British Antarctic, Faeroe Islands, Guernsey, Gibraltar, Greenland, Israel, Isle of Man, Jersey, Japan, St Kitts and Nevis, South Korea, Kuwait, Cayman Islands, Kazakhstan, St Lucia, Libya, Monaco, Yugoslavia: Montenegro, St Martin, Montserrat, Mauritius, Mexico, Malaysia, New Caledonia, New Zealand, Oman, Panama, French Polynesia, St Pierre and Miquelon, Pitcairn Island, Qatar, Yugoslavia: Serbia, Russia, Saudi Arabia, Seychelles, Singapore, St Helena / Ascension / Tristan da Cunha, San Marino, Turks and Caicos Islands, French Southern / Antarctic Territories, Trinidad and Tobago, United States of America, Uruguay, Vatican, Venezuela, British Virgin Islands, Wallis and Futuna Islands, Kosova, Mayotte, Former Serbia and Montegro, USSR (Ex), Yugoslavia (Ex)
MDC	Armenia, Azerbaijan, Bangladesh, Bolivia, Bhutan, Botswana, Belize, Congo, Cameroon, China (exc. Taiwan), Colombia, Cape Verde, Djibouti, Dominica, Dominican Republic, Algeria, Egypt, Fiji, Micronesia, Gabon, Grenada, Georgia, Ghana, Equatorial Guinea, Guatemala, Guyana, Honduras, Haiti, Indonesia, India, Iran, Jamaica, Jordan, Kenya, Kyrgyzstan, Cambodia / Kampuchea, Comoros, Laos, Lebanon, Sri Lanka, Morocco, Moldova, Madagascar, Marshall Islands, Burma / Myanmar, Mongolia, Mauritania, Maldives, Namibia, Nicaragua, Nepal, Peru, Papua New Guinea, Philippines, Pakistan, Palestine, Pacific Islands (inc Palau), Paraguay, Solomon Islands, Sudan, Senegal, Suriname, Sao Tome and Principe, El Salvador, Syria, Swaziland, Thailand, Tajikistan, Turkmenistan, Tunisia, Tonga, Taiwan (China), Tanzania, Ukraine, Uzbekistan, St Vincent and the Grenadines, Vietnam, Vanuatu, Samoa, Western, Yemen, South Africa
LDC	Afghanistan, Angola, Burkina Faso, Burundi, Benin, Congo Democratic Republic), Central African Republic, Ivory Coast, Western Sahara, Eritrea, Ethiopia, Gambia, Guinea, Guinea – Bissau, Iraq, Kiribati (and other Pacific Islands), Korea, North / DPR, Liberia, Lesotho, Mali, Malawi, Mozambique, Niger, Nigeria, Nauru, Rwanda, Sierra Leone, Somalia, Chad, Togo, East Timor, Tuvalu, Uganda, Zambia, Zimbabwe, Stateless

ESRC Centre for Population Change
Building 58, Room 2001
Faculty of Social and Human Sciences
University of Southampton
SO17 1BJ

T: +44 (0)2380 592579
E: cpc@soton.ac.uk
www.cpc.ac.uk

To subscribe to the CPC newsletter and keep up-to-date with research activity, news and events, please register online: www.cpc.ac.uk/newsletter

You can also follow CPC on Twitter, Facebook and Mendeley for our latest research and updates:

www.facebook.com/CPCpopulation

www.twitter.com/CPCpopulation

www.mendeley.com/groups/3241781/centre-for-population-change

The ESRC Centre for Population Change (CPC) is a joint initiative between the University of Southampton and a consortium of Scottish universities including St Andrews, Edinburgh, Stirling and Strathclyde, in partnership with the Office for National Statistics and National Records of Scotland.