Bayesian modelling of international migration with Labour Force Survey data

Arkadiusz Wiśniowski

Southampton Statistical Sciences Research Institute, University of Southampton

October 14, 2011

— DRAFT —

Abstract

A statistical method for estimating international migration flows based on information obtained from Labour Force Surveys (LFS) is developed in this study. The motivation for using LFS data is the general poor quality or absence of data provided by the national statistical institutes. A simple Bayesian model of flows between two countries is considered, assuming that migration can be measured in the LFS by both sending and receiving countries. The undercount in a particular destination-specific emigration flow is estimated by using the LFS sample of the corresponding receiving country. The model developed in this paper is applied to estimate recent migration flows between Poland and the United Kingdom.

1 Introduction

The purpose of this study is to propose a statistical framework for estimating international migration flows based on information obtained from Labour Force Surveys (LFS). The method can be used as a tool for validating and supplementing officially produced statistics.

The motivation for using LFS data is the general poor quality, inconsistency, or absence of data provided by national statistical offices in European countries. A simple Bayesian model of flows between two countries is considered, assuming that migration can be measured in the LFS by both the sending and receiving country. Further, the undercount in a particular destination-specific emigration flow is estimated by using the LFS sample of the corresponding receiving country. For illustration, the framework developed in this paper is applied to estimate recent migration flows between Poland and the United Kingdom.

Reliable and comprehensive statistics on international migration are required to analyse the reasons and consequences of the movements, as well as to design and implement fair and effective labour, economic and social policies. As noted by Bilsborrow et al. (1997, p. VI), lack of comprehensive data on migration leads to the analysis '[...] more in terms of impressions than in terms of impact' and to formulation of the policies on the 'shaky basis'.

2 Migration and migration measurement

Conceptually, international migration can be defined as a change of a place of a usual residence that occurs across a national border. A place of usual residence is defined as a place where a person lives, that is, spends his or her daily period of rest (United Nations, 1998). Willekens (1994, 2008) defines a place of usual residence as an address within administrative boundaries.

The number of migrants can be quantified in terms of stocks or flows. Stocks concern the size of the subpopulation of migrants in a given population at a certain moment in time. Flow of migrants is a number of movements of persons, who changed their country of usual residence during given period of time, e.g. a year or five years. Flows capture the dynamics of the migration processes (Bilsborrow et al., 1997).

In terms of measurement, flows can be measured by counting a number of movements, i.e. all changes of the place of usual residence within a period of time. In this case, if a person relocates a number of times, and all constitute migration events. The other way of quantifying flows is to compare usual places of residence at two points in time, e.g. one year apart. The first of the approaches is used in population registers or registers of foreigners. A second one can be used to measure migration flows by comparing the data from, e.g., two subsequent censuses. This approach is called a transition approach.

As far as migrant stocks are concerned, in both event and transition method the number of persons is measured (Rees and Woods, 1986, see Table 1). In the measurement of flows events are migration movements, transitions are different places of residence at two moments in time of a person. Hence, in transition approach persons are counted as migrants.

Table 1: Movements versus transitions approach

	Movements	**Transitions**
flows	migrations (events)	migrants (persons)
stocks	number of persons	number of persons

Source: Rees and Woods (1986)

Let M^{ij} denote all movements of all persons from country i to country j in a given period of time between points t and $t+1$. Let B_i and D_i denote all births and deaths respectively in country i between t and $t+1$. Then the population at time point $t+1$ in movement approach can be defined by the balance equation

$$P^i(t+1) = P^i(t) + B^i - D^i + M^{Oi} - M^{iO}, \tag{1}$$

where superscript O (from 'outside') denotes the the rest of the world.

In terms of transitions the same population at time $t+1$ can be defined in the following way. Let K^{ij}_{t+1} denote the number of persons present in country i at time point t and in country j at time point $t+1$, K^{Bji}_{t+1} denote persons born in country j that are counted in country i at time point $t+1$. A dot $(\cdot)$ denotes a summation over the index. Then the balance equation is

$$P^i(t+1) = K^{\cdot i}_{t+1} = K^{ii}_{t+1} + \sum_{j \neq i} K^{ji}_{t+1} + K^{Bii}_{t+1} + \sum_{j \neq i} K^{Bji}_{t+1}. \qquad (2)$$

2.1 Data collection in Poland

Since 2006, the most important source of information on migrants is official statistics register of permanent migration (*Powszechny Elektroniczny System Ewidencji Ludności*, Universal Electronic System for Registration of the Population, PESEL), which allows for measurement of both flows and stocks of migrants. Information on temporary migrants for more than 3 months is also collected, however, not all immigrants register and only a small fraction of emigrants report their departure (Nowak et al., 2008). The register gathers information about the sex, age and nationality of immigrants, as well as the origin or destination countries.

Another source of data on migrants is a register kept by Office For Foreigners 'System POBYT'. This register gathers information on foreigners (residents of the EU) who apply for permanent or temporal stay permits, and asylum seekers. Information on migrant characteristics, such as sex and nationality, is also collected.

Data on work permits issued to immigrants is collected by the Ministry of Labour and Social Policy. However, not all immigrants are required to obtain such permits. The Ministry of Interior and Administration gathers data on repatriates, persons who obtained Polish citizenship and the whole population of Poland by nationality, age and sex. Both public and private higher level education institutions gather gather data about their students and graduates by sex, age and nationality.

Another source of detailed data on migrants are censuses, which have been conducted every ten years since 1921. The most interesting rounds are a 2002 and the forthcoming 2011 round Census. However, in this paper we will not discuss them in detail.

The Polish Labour Force Survey, (*Badanie Aktywności Ekonomicznej Ludności*, BAEL) is a survey-based source of data on migrants. Information on nationality, country of birth and a place of residence one year before the survey is gathered for all persons in the household. More information on the LFS in Poland is presented in the next section.

Finally, in the European Union Statistics on Income and Living Conditions, information is gathered about all persons in sampled households. Available data concern sex, country of birth, nationality, reason for absence, time of being abroad, country of residence abroad. The sample size is too small for measuring detailed characteristics of the migrants and the results obtained in this survey are not disseminated (Nowak et al., 2007).

All data sources mentioned above report to the Central Statistical Office in Poland, but not all of the processed data are disseminated.

2.2 Data collection in the UK

A single and comprehensive data source on migration flows or stocks in the United Kingdom does not exist (Ker et al., 2009). All disseminated data are obtained from samples that are designed for different purposes. The data are collected and disseminated by the Office for National Statistics (ONS).

Data on flows of migrants are obtained by combining the results of the International Passenger Survey (IPS), which surveys people leaving and entering the United Kingdom, with data from the Home Office and Irish Central Statistics Office (until 2007). It produces estimates of flows called Long-Term International Migration (LTIM). Size of the IPS survey is about 230,000 (2008) and response is around 83%. In 2008 2.2% (5,117) of interviewees were actual migrants (Ker et al., 2009). Collected characteristics include origin or destination country, nationality, country of birth or reason for migration. Survey is based on intentions of migrants, which may not necessarily reflect their actual actions. The numbers are corrected for actual length of stay of migrants at the later stage by using estimation techniques and the data obtained in the previous waves of the survey. Home Office adjusts the final numbers adding the asylum seekers.

The main source of data on stocks of migrants is the Labour Force Survey supplemented by the Annual Population Survey (APS). The APS is based on the LFS sample, with additional 'boost' surveys designed to ensure representation of each area in the UK (Ker et al., 2009).

Administrative sources of data on migrants are National Insurance numbers (NINos), Work Registration Scheme (WRS), which captures mainly workers from Eastern and Central Europe. Some reconciliation efforts have been undertaken by ONS, which combined data from the IPS with numbers provided by the WRS, NINos and Patient Register Data System (PRDS, report 'Reconciliation of ONS estimates: Comparisons of combined IPS (long and short term migration) estimates with administrative data sources', ??).

2.3 Quality and availability of international migration data

There are three distinctive problems concerning the quality of the statistical data on migration: availability, reliability and comparability (Nowok et al., 2006). Availability of the data on migration depends on the specific features of the particular data-collection systems.

The data are said to be reliable if they comply with the definition adopted by the national data-collecting system. Poor reliability of the statistical data on international migration flows is a well known fact (cf. Bilsborrow et al., 1997; Nowok et al., 2006, Fassmann, 2009, de Beer et al., 2010). The main reason of such a situation is the under-registration of the migrants, especially in the collection systems based on self-declarations. This undercount is thought to be more severe in the case of emigrants, who usually have less incentives to deregister from the system than immigrants, who, after registration, may gain access to certain benefits, such as health insurance, pension schemes and social benefits.

The data that are unreliable are often incomplete and or inconsistent, thus preventing any comparisons between countries (see Nowok 2010). If we consider a number of migrants between two countries, this number can be reported by the sending and by the receiving country. These two numbers never match each other, which results from the different definitions of a migrant adopted by countries and different data collection mechanisms. For instance, the flow from Poland to the United Kingdom in 2006 reported by Polish GUS is about 18,000 people, whilst the number reported by the UK's IPS is 60,000.

As it is noted by Kupiszewska and Nowok (2008): 'In recent decades there have been many efforts to harmonise migration statistics [...]. However, the results are far from satisfactory'. De Beer et al. note even that '[...] the situation today in terms of migration definitions and measurement is not much better than it was [...] 20 years ago' (de Beer et al., 2010).

The sources of the lack of harmonisation of the data lie in the procedures of data collection adopted by particular countries. There are three stages at which the differences between countries are found:

1. collection of raw data in the primary data source,

2. production of the statistics,

3. dissemination of the data (Kupiszewska and Nowok, 2006: p.).

In the first stage the most important role plays the country's legislation and resulting from it definition of a migrant and of a migration event. If the concept of the usual residence is used it is most often interpreted in legal (*de jure*) terms. It means that it is required for a person to fulfill certain criteria in order to become a resident and be considered a migrant. One of these criteria is the duration of stay, e.g. no time limit is adopted in Germany, 12 months in Sweden or permanent stay in Poland. Duration of stay may also be intended or actual. The usage of the intended stay approach leads to errors in the data, actual length of stay causes a systematic delay in the production of statistics.

In country-specific systems, different subpopulations may be covered. Usually, official statistics cover only legal migrants, some countries (e.g. France or Portugal) do not consider nationals as migrants. Data collection systems may be inadequate for providing the required level of detail of the data, such as estimates produced by the passenger survey in the United Kingdom are affected by high estimation errors. All these above mentioned differences between national data-collection systems lead to discrepancies in official statistics on international migration. These discrepancies are exposed best when comparing the figures for immigration in the destination country and emigration in the origin country.

Recently, there have been some attempts on improving the comparability of the data on international migration in European Communities. Regulation no. 862/2007 of the European Parliament and of the Council of Europe from July 2007 provides common definitions and guidance for the collection of the data on migration (Official Journal of the European Union, 2007). It allows using well documented and scientifically based estimation methods for processing of the data.

Lemaitre (2005) proposed the idea of using reason for movement to harmonise statistics on migration. It was partially implemented in the International Migration Outlook 2006 (OECD, 2006), a continuation of the SOPEMI reports. These statistics are called 'standardised' and concern permanent immigration of foreigners. Permanent immigration is an immigration of a person with a residence permit which is either permanent or more or less renewable (OECD, 2008).

One of the first efforts to provide methodology for the harmonisation of the international data on migration were made within the MIgration MOdeling for Statistical Analysis (MIMOSA) project, funded by Eurostat and coordinated by the Netherlands Interdisciplinary Demographic Institute. The result of the project was the first complete, harmonised and consistent set of estimates of international migration flows between EU and European Free Trade Association (EFTA) countries, see de Beer et al. (2010) basing on the works by Poulain (1993, 1999) and Poulain and Dal (2008). All estimates were based on the mean square error minimising algorithm. However, the method is incapable of providing any measures of uncertainty of the estimates.

On the basis of the MIMOSA, the follow up project funded by NORFACE, Integrated Modelling of European Migration (IMEM), has been undertaken by the teams at the University of Southampton, Oslo and at NIDI. In the this project, a Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows is proposed (Raymer et al., 2010). The focus is on estimating recent international migration flows between countries in the European Union (EU) and EFTA, using data primarily collected by Eurostat and other national and international institutions. The methodology is integrated and capable of providing a synthetic data base with measures of uncertainty for international migration flows and other model parameters.

3 Labour Force Survey

Labour Force Surveys (LFS) are quarterly surveys (since 2005) carried out in EU, EFTA and candidate countries. They are aimed at providing insight into characteristics of the labour market in Europe.

LFS was never intended to measure mobility of people. The questionnaire, however, includes questions which makes estimation of both stocks of immigrants and the international flows of migrants possible. The variables that allow for identification of stocks are nationality and country or place of birth. The question about the country of residence one year before the survey identifies the flows in a transition approach, such as in censuses, by comparing the place of residence at two different points in time.

The transition approach in measurement of flows means that all movements in between the compared points in time are neglected. The information on residence changes can be revealed by the surviving immigrant when he or she is present in the household. An emigrant can be identified by the other members of the household, in which this emigrant used to reside one year before the interview, providing they exist and are present at the time of the interview. If the whole household emigrated, the identification is impossible.

Migration in the LFS is a very rare event, even if estimation of migrant stocks is of interest. The UK's International Passenger Survey, which is aimed at measuring the flows of migrants to and from the United Kingdom, has samples a few times larger than the ones in the LFS and they are still considered to small to capture the country of origin or destination structure (Nowok, 2006). Rareness of migrants in the sample can be treated as a sampling error, that is an error occurring when inference about a population is made using data from a sample. Sampling error affects the accuracy of the estimates.

In the LFS, apart from the rareness of migrants in the samples, there are at least four other problems which may possibly influence the analysis of the results. These are non-response, one reason of which is refusal of answer, using proxies and an undercount. Non-response can occur when a respondent cannot be contacted or refuses to answer. Non-response can be a source of bias of the estimates, if persons who do not respond differ from those who do. In the case of migrants non-responsiveness can be expected, especially for illegal migrants or those who do not speak the language of the country. The level of non-response in the Polish LFS is about 25%, in the UK it is 30%. Refusal is one of the reasons of non-response, but it can be potentially strongly correlated with a migration event. Refusal rate in Poland is 59% of non-response, for the UK it is 72%. A proxy is taking an answer to the questions given by another related adult, who is a member of a household, if a respondent is not available. In the case of migrants it can be expected that information on those who do not speak the language of the LFS country may be obtained from proxy respondents. In Poland proxies constitute around 41% of all answers, in the UK 36% (LFS User Guide, 2009). Undercount of migrants, as it was mentioned before, concerns the respondents who leave a country with the whole household (sample escape). It is impossible to learn about undercount from a survey for a single country. However, the idea of this paper is to use complimentary information on migrants that can be obtained from surveys conducted in several countries. This means that the whole households migrating from country A to country B, which are not captured by the LFS in country A as emigrating, are included in the sampling population and can be captured in the LFS in country B as immigrating ones.

3.1 LFS in selected countries

Despite common EU regulations and guidelines concerning the LFS construction, the design of the LFS differs across countries. Differences concern the sampling frame, stratification, rotation, final sampling unit.

3.1.1 LFS in Poland – BAEL

BAEL carried out in Poland is voluntary and covers the population aged 15 or more years old. It is based on 2002 Census balances and National Official Register of Territorial Division of the Country (Domestic Territorial Division Register).

There are four elementary samples for each quarter, which comprise of 43200 (2008) persons altogether, that is about 0.1% of the total population. All four quarterly samples are divided to 13 weekly elementary samples. The Primary sampling units (PSU) are census clusters in towns and enumeration districts in rural areas, second stage sampling units constitute dwellings. The final sampling units are households. BAEL does not cover the subpopulation living in the institutional households, such as lodging-houses for employees, student halls of residence, boarding-schools, army barracks, social welfare homes.

In BAEL there is a rotation scheme 2-(2)-2, which means that a household is interviewed in two consecutive quarters, then it has two quarters break and then is interviewed again in the next two quarters (see Table 2). In this way a household stays in the survey for one and a half years. The rotation scheme is presented in the table below. Stratification variables used in BEAL are 1) voivodships (large administrative areas in Poland, pol. *województwo*), 2) urban/rural areas and 3) strata within voivodships (two to four, depending on the voivodship). Poststratification includes the variables sex, age, rural/urban areas and six categories of place of residence.

Table 2: Sampling scheme in the Polish LFS

No.	quarters							
próby	3	4	1	2	3	4	1	2
24	X							
25	X	X						
26	-	X	X					
27	-	-	X	X				
28	X	-	-	X	X			
29	X	X	-	-	X	X		
30		X	X	-	-	X	X	
31			X	X	-	-	X	X
32				X	X	-	-	X
33					X	X	-	-
34						X	X	-
35							X	X

3.1.2 LFS in the United Kingdom

In the United Kingdom, the LFS is voluntary and covers population aged 16 or more. It is based on the Postcode Address File or telephone directory (especially in Scotland).

There are five elementary samples which consist of 52100 dwellings (2008), which represent about 0.1% of the total population. Similarly as in Polish BAEL, the samples are divided into 13 weekly elementary samples. Sampling units are addresses or phone numbers, thus the interviews are carried out face-to-face or over the phone. Institutional households (e.g. students in halls of residence and patients of NHS, prisoners and military servants) are excluded from the sample.

The rotation scheme in the LFS is 5-, that is a household is interviewed for five consecutive quarters (see the Table 3). Private addresses are used as a stratum for sampling. For post-stratification sex, age and regions are used. It should be noted that the question about country of residence one year before an interview is asked in the second (spring) quarter only, which precludes more thorough identification of migrants in the samples.

Table 3: Sampling scheme in the British LFS

No.	quarters							
próby	3	4	1	2	3	4	1	2
1	X							
2	X	X						
3	X	X	X					
4	X	X	X	X				
5	X	X	X	X	X			
6		X	X	X	X	X		
7			X	X	X	X	X	
8				X	X	X	X	X
9					X	X	X	X
10						X	X	X
11							X	X
12								X

3.2 Attempts on using LFS to measure migration

Migration using stocks based on the BAEL have been included in the OECD SOPEMI reports, since 1996 disseminated by Centre for Migration Research (CMR) at the Warsaw University, see e.g. Okólski (1996) or Kępińska (2007).

In 2007 Centre for Migration Research (CMR) at the Warsaw University built a database based on the BAEL from years 1999-2008 (first quarter). The database was used to assess the information about the returning migrants after two or three months of staying abroad. The database contains 6338 emigrants and 542 returning migrants for the whole observed period. It should be noted that the duration of stay criterion refers more to a short-term migration rather than long-term according to the UN recommendation (1998). The largest group of migrants in the database are persons who went to the UK, 24.6%, the second most popular country is Germany, with 22.6%. Most of the returning migrants come from these countries, 33.4% from Germany and 15.9% from the UK.

Anacka (2008) used the CMR database to compare characteristics of Polish migrants before and after 1 May 2004, that is, the day of the accession of Poland to the European Union. By means of migration selectivity index and a simple logit model, she investigated the propensity for migrating. The categories used to compare migration patterns include the migrant characteristics (age, sex, education, size of the place of origin) and of household characteristics (main source of income, number of persons in the household).

LFS is used in all EU countries to acquire information on migration movements by introducing a special module concerning migration in the 2008 LFS, based on the regulation EC 102/2007 from 2 February 2007 (Official Journal of the European Union, 2007). In Poland, the special LFS module was carried out in the second quarter of 2008. Some of the questions were optional. A specific part of the module in Poland concerned returning migrants, that is, persons who ever stayed abroad for more than three months. In the survey, a question about country of birth of parents was included in order to separate the second generation of immigrants. The questions of the module were incorporated in the LFS standard survey, hence the results should be treated as tentative. The sample included 42700 persons in total.

The LFS special module on migration was combined with the survey 'Imigranci w Polsce' (Immigrants in Poland) targeted at the institutional households, as they are not incorporated in the LFS (US, 2008, Informacja o badaniach zasobów imigracyjnych w Polsce w 2008 r.). The survey encompassed 10242 immigrants from 131 countries. From the institutional households there were excluded hospitals and other medical institutions, orphanages and other educational institutions, monasteries and asylum seeker centres.

The special module in the UK LFS was carried out by ONS (2009). It was aimed at more in-depth study of the migrants situation in the UK, such as reason for coming, social and labour market integration and adaptation or qualifications held.

LFS data, combined with the ad hoc Eurostat module from spring 2008, have been extensively used to analyse the employment of foreign workers in the UK. The analysis concerned period of arrival, male and female labour market participation, earnings and the outcomes from the special module. Introduction of the question about a month of arrival to the UK allows for an analysis of the patterns of the short-term migrants (or those who arrived to the UK recently) in the labour market (Ker et al., 2009).

Ker et al. (2009) list all definitional differences between the sources of data on migration in the United Kingdom. They compare the LFS and Annual Population Survey (APS), which is used for statistics on stocks of migrants by nationality and country of birth, with International Passenger Survey (IPS), which is used for production statistics on flows (LTIM, see section 2.1.3). While the IPS uses the UN (1998) recommended definition of a long-term migrant, in the LFS a migrant is defined as not born in the UK or not having British nationality (country of birth is preferred as it cannot change, while nationality can, ONS, 2009). Duration of stay is not used, however LFS asks for the date of arrival to the UK. LFS excludes students and asylum seekers living in the communal establishments, while both groups are captured by the IPS, with asylum seekers being adjusted by the Home Office.

Drinkwater et al. (2006) use British LFS for analysis the labour market outcomes in terms of the relative earnings of migrants. They focus on migrants from Poland and the other seven Eastern and Central European countries (A8) before and after the accession in May 2004. For comparison purposes the migrants from other European countries, English speaking countries and other countries are included in the study. Their dataset covers the period from Autumn 2001 to Summer 2006. In the sample, the number of immigrants from Poland who came before year 2000 is 234, 169 who came between 2000 and 2003 and 259 of those who came after the enlargement. The duration of stay criterion is not relevant. They also find that the characteristics of the migrants in the LFS sample are similar to those reported in the Worker Registration Scheme (WRS). The main finding of their study is that many migrants from A8 group are employed in very low paying jobs despite having relatively high levels of education.

Martí and Ródenas (2007) evaluate the quality and potential comparability of the migration statistics on stocks and flows based on the LFS surveys in the 15 'old' EU countries (EU-15). They identify the possible statistical problems that hinder the production of statistics especially on migration flows, namely bias and imprecision (accuracy). In the EU-15 countries, LFS can be used mainly for statistics on stocks of active foreign population. Some countries, such as the United Kingdom, Spain, Belgium, France and Portugal refer to it due to lack of other sources. Martí and Ródenas find out that estimates of flows based on the LFS samples are much less accurate than those based on the registers or censuses. In general, flows are underestimated apart from the cases of Austria (comparing to register-based reported numbers), France and Portugal (comparing to census-based reported numbers), whereas for stocks the discrepancies are significantly smaller, yet the results differ between countries. They suggest it is the result of the specific national sample design of each country. However, the criteria used by statistical offices for producing statistics on migrants and their potential influence on the differences between them and the LFS estimates seem to be neglected in the study.

Martí and Ródenas (2007) claim the accuracy of the estimates depends on the size of the sample comparing to the domain, heterogeneity of the studied variable and efficiency of the stratification used. Sampling the households or dwellings as the primary sampling units, instead of persons, may also result in high errors in the estimates. Stratification (or poststratification) may reduce the sampling error if the stratum is correlated with the sampled object (migrants). For instance, in Sweden nationality is used for stratification, in Austria, Germany, the Netherlands and Luxembourg it is used for poststratification and hence it may help to reduce error in estimating the stock of nonnationals. The bias in the estimates results from suitability of the sampling frame (such as sampling only private households), its updating (based on registers or on census) and nonresponse (due to not residing in the household or dwelling or refusal).

Martí and Ródenas (2007) point out the problem of *answer impossible*, that decreases accuracy in estimating flows of migrants. The problem arises from the time criterion used in the question about the place of residence one year before the interview. Rotation schemes adopted in countries, as well as rest quarters between interviews, imply different durations of participation in the survey. When an immigrant stays in a sample for longer than one year, at some point he or she is never able to give a positive answer concerning his or her migration (change of the place of residence). The proportion of the sampling units that stays in a sample a year after first survey that is less than 50% is found only in four countries (Belgium, Denmark, Luxembourg and the Netherlands).

4 Model of flows - conceptual framework

In this section, a conceptual framework of a statistical model is presented, which serves as a tool for measuring migration flows using data from the LFS. The adopted perspective for estimation of the model parameters is Bayesian. It provides a coherent framework for making statistical inference and allows to analyse scarce or poor quality data, with the possibility of introducing the expert knowledge in by means of subjective a priori distributions (sf. Bijak, 2010, Bijak and Wiśniowski, 2010, Raymer et al., 2010).

The model is based on the assumption that emigration is measured with bias in sending countries. The main source of this bias is emigration of the whole households. For handling this bias a separate parameter is introduced. Its purpose is to capture the undercount and learn about it from complimentary information from both sending and receiving countries.

At first we consider $r = 2$ countries and category 0 as the 'rest of the world'. K_t^{ij} denotes the 'true' count of persons who are in country j on the 1st January of year t and were in country i on the 1st January $t-1$ year. Then the 1st January population N_t^1 in a year t in country $i = 1$ is given by the balance equations

$$\begin{aligned} N_t^1 &= K_t^{\cdot 1} = K_t^{11} + (K_t^{21} + K_t^{01}) + B_t^1 && (3) \\ &= K_{t+1}^{1\cdot} = K_{t+1}^{11} + (K_{t+1}^{12} + K_{t+1}^{10}) + D_{t+1}^1. && (4) \end{aligned}$$

In country $i = 2$ the population is given by

$$\begin{aligned} N_t^2 &= K_t^{\cdot 2} = K_t^{22} + (K_t^{12} + K_t^{02}) + B_t^2 && (5) \\ &= K_{t+1}^{2\cdot} = K_{t+1}^{22} + (K_{t+1}^{21} + K_{t+1}^{20}) + D_{t+1}^2. && (6) \end{aligned}$$

N_t^i - population on 1 January t year, $B_t^i \equiv K_t^{Bi}$, $D_t^i \equiv K_t^{iD}$ - born and dead between $t-1$ and t, a dot (.) denotes summation over index.

Using the balance equations the model of flows is derived assuming the ratio of the true migrants to the population is proportional to the observed number of migrants in the LFS sample to the sample size. It has to be noted that the sample does not reflect the whole population, e.g. the lack of institutional households in the sample.

We observe that the true flows K_t^{12} are the same in the balance equation for year t of the country 2 (then they are immigration) or in the balance equation for year $t-1$ for country 1 (then they are emigration). The measured sample data are denoted by small letters: k_t^{ijm}, where $m = \{S, R\}$ indicates whether the measurement comes from the sending or receiving country. n_t^j denotes the sample size in the given LFS study.

We assume that the true initial risk (which we interpret as a probability) p_t^{ijS} is a migration average risk measured as a ratio of the counts of migrants from country i to j, K_t^{ij}, to the population at risk, that is N_{t-1}^i,

$$p_t^{ijS} = \frac{K_t^{ij}}{N_{t-1}^i}. \tag{7}$$

A ratio of the true number of migrants to the population size of the receiving country can interpreted as a probability of being an immigrant, that is

$$p_t^{ijR} = \frac{K_t^{ij}}{N_t^j}. \tag{8}$$

Now we can define a binomial model for flows between two countries. The probabilities p_t^{12S} and p_t^{21S} are probabilities of being observed as an emigrant from country 1 to 2 and from country 2 to 1, respectively, as well as the probabilities of being observed as an immigrant in country 2 and 1, respectively, that is p_t^{12R} and p_t^{21R}. Then the relation between the two probabilities is

$$p_t^{ijR} = \frac{N_{t-1}^i}{N_t^j} p_t^{ijS}. \tag{9}$$

Assuming $N_t^{ij} \equiv N_{t-1}^i / N_t^j$, we can write the model as

$$k_t^{12R} \sim \mathcal{B}\left(n_t^2,\ N_t^{12} p_t^{12S}\right), \qquad k_t^{12S} \sim \mathcal{B}\left(n_{t-1}^1,\ \lambda p_t^{12S}\right) \tag{10}$$

$$N_t^{ij} = \frac{N_{t-1}^i}{N_t^j}. \tag{11}$$

In the model 10 there is an assumption that there is a systematic undercount of migrants in the sending country (by not counting whole emigrating households). It is measured by means of a positive parameter $\lambda \in (0, 1)$, possibly time- and sending-country-specific and depending on some covariates, i.e.

$$\lambda_{(t)}^{ij} = \prod_{k=1}^{K} \phi_k^{Y_t^{ijk}}.$$

The second considered model is based on the assumption that the LFS counts of migrants, k_{ij}, are Poisson distributed. We assume that the average risk of migration is proportional to the LFS sample counterpart of this risk, but it is measured with a log-normal error resulting from the sampling scheme. The errors and their precisions are country-specific. It is also assumed that emigration is measured with a bias denoted again by the parameter $\lambda \in (0,1)$. Using equations 7 and 8 we can write equations for K_t^{12}

$$\frac{K_t^{12}}{N_t^2} = \frac{k_t^{12R}}{n_t^2} e^{\xi_t^{12R}}, \qquad \frac{K_t^{12}}{N_{t-1}^1} = \frac{k_t^{12S}}{n_{t-1}^1} e^{\xi_t^{12S}} \frac{1}{\lambda}$$

$$k_t^{ijR} \sim \mathcal{P}(\mu_t^{ijR}), \qquad k_t^{ijS} \sim \mathcal{P}(\lambda \mu_t^{ijS}), \qquad \xi_t^{ijk} \sim \mathcal{N}(0, \tau^k)$$

Comparing to the binomial model, the Poisson model is much more flexible in capturing the uncertainty of the flow estimates by means of the log-normal error term.

Both models have been thoroughly analysed using a simulated set of data and applying the MCMC algorithms in order to estimate the posterior densities of the model parameters (see e.g. Robert and Casella, 2004). For the brevity of this paper the results of this statistical exercise are not presented.

5 Estimation of flows between Poland and the United Kingdom

In this section we apply the methodology described in previous ones to estimate migration flows from Poland to the United Kingdom. The results are preliminary and are based on two years of observations only. First, we analyse two models, the binomial and the Poisson. Then we perform sensitivity analysis with respect to various assumptions about the model parameters. Finally, we discuss the preliminary results.

5.1 Model of flows between Poland and the United Kingdom

In Tables 4 and 5 the counts of migrants in the Polish and British LFS, respective sample sizes and population sizes (aged 15 and older for Poland and 16 and older for the UK) are presented. In general, the number of immigrants to a country is derived from the answers to the question 'where have you been one year ago'. The number of emigrants is calculated by comparing variables providing information on whether a person was in the country at the moment of the survey or was abroad for more than 2 or 3 months, with an information on the country in which the absent person was staying at the time of the survey. For the moment, due to the characteristics of the data available at the time of the study, such identification is possible only for the emigrants from Poland to the United Kingdom. Moreover, the data from the Polish LFS are available only since 2007. Hence, the data allow estimating only one way flows from Poland to the UK in years 2007 and 2008.

Table 4: Data on migration flows in Polish and British LFS

Year	Polish LFS flows		British LFS flows	
	from UK to PL	from PL to UK	from PL to UK	from UK to PL
	k_{21R}	k_{12S}	k_{12R}	k_{21S}
2007	12	66	153	NA
2008	17	42	93	NA

Table 5: Data on population and LFS sample size in Polish and British LFS

Year	LFS sample size		Population size (age 15+)	
	PL	UK	PL	UK
	n_1	n_2	N_1	N_2
2007	27,680	123,715	32,103,119	49,258,093
2008	25,955	122,049	32,214,763	49,681,273

First we consider the binomial model. By K^{21} we denote the number of migrants and by p^{21} the rate of emigration from Poland to the United Kingdom, averaged over 2007 and 2008. Parameter λ measures the undercount of emigrants in the Polish data and is also constant over time. For the Bayesian analysis the following prior parameters are assumed:

$$\lambda \sim \mathcal{U}[0,1]$$

for the undercount parameter ($\mathcal{U}$ denotes uniform distriution) and

$$p^{ij} \sim \mathcal{B}(1,1000)$$

for the emigration from Poland rate ($\mathcal{B}$ denotes Beta distriution). The prior for the undercount implies the same probability for each of the values from between zero and one, that is a total lack of knowledge on the potential size of the undercount. The second prior implies some subjective knowledge about the rate of migration. The expected value of it is about 0.001, which means that, on average, one person per thousand emigrates from Poland to the United Kingdom. If we translate this into number of migrants, then with 95% probability it would be between 760 and 120,000. We believe that this range is realistic and wide enough to cover the possible outcomes, especially if we know that the total reported flows to the UK from all the countries are about 500,000 people.

The results of applying the binomial model are presented in Table 6. We observe that the posterior mean of the rate of migration is 1.5 person per thousand. This leads to the number of migrants aged 16 years old or more from Poland to the United Kingdom to be around 48,600 persons per year, averaged for years 2007 and 2008. With 95% probability we can say that this annual average lies between 42,800 and 54,830. The undercount indicated by the data is 0.6 with standard deviation of 0.07, which means that in the Polish sample only 60% of the emigrants are observed.

Secondly, the data from the Polish and British LFSs are analysed by using a Poisson

Table 6: Posterior characteristics of binomial model parameters

par	mean	SD	p2.5%	median	p97.5%
λ	0.598	(0.071)	0.471	0.593	0.752
p_{21}	1.52	(0.10)	1.33	1.51	1.71
K_{21}	48,660	(3,088)	42,800	48,610	54,830

model. The notation is analogous to the binomial model. The prior density assumed for K^{21} flow is constructed as

$$K^{21} = p^{12}N^1, \quad p^{12} \sim \mathcal{B}(1, 1000),$$

where N^1 is a population size in Poland ('population at risk', i.e. the population from which the migrants leave). This prior implies the same values for the number of migrants as in the binomial model, hence it allows for comparisons between the models under consideration. For the undercount parameter λ we assume the same uniform prior on a range between zero and one.

Prior density assumed for the precision parameters τ is Gamma $\Gamma(1,1)$. Since this choice is arbitrary, we explore the other combinations of the hyperparameters, such us $\Gamma(10,10)$ and $\Gamma(100,100)$. We also check how an assumption about the precision of the Polish LFS equal to the British one influences the results.

In Table 7 the outcome of the simulation is presented. We observe that the mean number of migrants is 59,140 people annually, averaged for years 2007 and 2008. The 95% probability interval is wider than the one resulting from the binomial model, that is the flow lies between 22,400 and 120,800. The a posteriori mean of the undercount is 0.7 with standard deviation 0.2, which is about three times larger than the posterior uncertainty in the binomial model.

Table 7: Posterior characteristics of Poisson model parameters

par	mean	SD	p2.5%	median	97.5%
λ	0.707	(0.203)	0.281	0.740	0.988
τ^S	1.41	(1.12)	0.14	1.13	4.42
τ^R	1.68	(1.25)	0.21	1.35	4.94
K_{21}	59,140	(25,470)	22,390	54,610	120,800

The additional uncertainty stems from the assumption that, on top of the Poisson variability, the observations from the LFS have a symmetric measurement error, reflected by the normal random term in equations for emi- and immigration counts. Prior assumptions about the precision of these error terms may have an influence on the uncertainty assessment of the migration flows. Hence it requires a sensitivity analysis with respect the parameterisation of the prior.

First we investigate the sensitivity to the assumption about the hyperparameters of the prior. We consider two alternative specifications, namely $\Gamma(10,10)$ and $\Gamma(100,100)$. Both of them have the same expected value with varying variance. In Table 8 the results of the analysis are presented. We observe, that despite characteristics of the error terms' precision change, the posterior mean and quantiles of the parameters λ and the number of migrants K remain almost the same.

Table 8: Sensitivity analysis of the Poisson model - prior for precision

par	mean	SD	p2.5%	median	97.5%
$\Gamma(10,10)$					
λ	0.704	(0.203)	0.274	0.735	0.989
τ^S	1.11	(0.32)	0.56	1.08	1.83
τ^R	1.00	(0.32)	0.49	0.96	1.74
K_{21}	59,470	(26,560)	21,710	54,700	124,100
$\Gamma(100,100)$					
λ	0.703	(0.213)	0.248	0.743	0.988
τ^S	1.01	(0.10)	0.83	1.01	1.22
τ^R	1.00	(0.10)	0.82	0.99	1.20
K_{21}	59,020	(26,050)	22,620	54,320	124,700

In the next step we analyse the sensitivity of the results to the assumption that the precisions of measurement errors in both sending and receiving countries are equal. This may seem to be plausible as we have the true flow K in both equations, hence the error terms reflect the uncertainty concerning the measurement of the same quantity, but in two different data collection systems. Another reason for investigating sensitivity to this assumption is a potential weak identifiability of them due to the same reason, as well as the scarce number of observations per parameter - with two precision parameters, we only have two observations per each in the estimation procedure.

In Table 9 we present the results of the analysis. We observe no significant influence on the a posteriori characteristics of the migration flows. The undercount parameter seems to be slightly larger for the case with the prior $\Gamma(1,1)$ and smaller for the $\Gamma(100,100)$ specification, but its uncertainty remains large and unchanged. Hence we draw a tentative conclusion that the precision of both counts of migrants in Poland and the UK can be the same.

5.2 Discussion of the results

The estimated number of migrants from Poland to the UK is about 60,000 people. This is an annual average for years 2007 and 2008. The number of migrants reported by the Polish register is 9,165 in 2007 and 22,352 in 2008. The British data for these years, coming from the IPS, are unavailable in the Eurostat database (as of 18 August 2011). The last reported value for flows in 2006 is 58,486 people. We need to keep in mind that all the officially reported data concern the whole population and the intentions of the potential migrants. In the case of the estimates based on the LFS data presented above, the migrants are aged 16 years old or more and they have actually arrived in the UK within the last year.

Table 9: Sensitivity analysis of the Poisson model - equal precisions assumption

par	mean	SD	p2.5%	median	97.5%
$\Gamma(1,1)$					
λ	0.727	(0.191)	0.310	0.762	0.988
τ	1.91	(1.27)	0.31	1.62	5.07
K_{21}	59,430	(24,060)	23,100	55,820	115,500
$\Gamma(10,10)$					
λ	0.704	(0.203)	0.274	0.735	0.989
τ	1.11	(0.32)	0.56	1.08	1.83
K_{21}	59,470	(26,560)	21,710	54,700	124,100
$\Gamma(100,100)$					
λ	0.687	(0.207)	0.277	0.711	0.987
τ	1.01	(0.10)	0.83	1.01	1.21
K_{21}	60,040	(27,060)	22,050	55,690	125,300

The LFS estimate of the flow has some obvious shortcomings. First of all, it provides only an annual average, thus it shows no dynamics of the flows over time. However, model specification allows to capture the time effects with the data becoming available over time. Secondly, the estimate relies on the population at risk, which, in this case, is the population of Poland as reported by the registers. But if we believe that the number of people going from Poland to the UK is different from the reported one, then it obviously affects the underlying population size. The intensified outflow of Polish migrants has been observed since the accession to the European Union in 2004. The answer to the question about the ‘true’ population size after the EU enlargement lies in the forthcoming results of the 2011 Census round, in which the total number of people actually residing in a country is counted. The second problem with the population at risk used for the estimates lies in the LFS sampling scheme. As described in previous sections, Polish LFS is not carried out in the institutional households. Hence, the reference population should be lower than the officially reported but it is difficult to estimate the total number of people residing in these households. Moreover, stratification of the sampling units may bring about further changes to the population, as some regions (Primary Sampling Units) have comparatively larger samples than the other. This bias can potentially be adjusted by using weights for each of the PSU. The third flaw of the modelling framework is neglecting the information about the non-response and refusals of answering. These variables may carry information about the undercount of the migrants, measured in the receiving country. However, this information is aggregated and concerns respondents who can be both immi- and non-migrants. In order to take the non-response percentage into account, the second stage of the estimation can be considered. The non-respondents could be distributed proportionally to the immigration (or emigration) rate obtained in the first stage. Then the new number of migrants could be computed and the whole estimation procedure repeated. This would lead to the higher estimate of the flows.

Summarising, bearing in mind all the caveats discussed above, we believe that the estimated number of people migrating from Poland to the United Kingdom proves that the method is capable of providing a supplementary information on the migration flows between countries. The drawback of the preliminary application of the method is that it encompassed only one way flow between the two countries and the data available at the time of the analysis were available only for two years. However, extending the model to include flows within a system of countries will yield the more precise estimates of the flows and the other model parameters. It also provides a room for including the additional covariates in the model.

References

[1] Anacka M (2008) Najnowsze migracje z Polski w świetle danych Badania Aktywności Ekonomicznej Ludności. CMR Working Papers, No. 36/94, Warszawa

[2] Bijak J (2011) *Forecasting international migration in Europe: A Bayesian view.* Dordrecht: Springer.

[3] Bijak J and Wiśniowski A (2010) Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 173(4):775-796.

[4] Bilsborrow RE, Graeme H, Amarjit SO, and Zlotnik H (1997) *International migration statistics: Guidelines for improving data collection systems.* Geneva: International Labour Office.

[5] 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* 26:459-481.

[6] Drinkwater S, Eade J, Garapich M (2006) Poles Apart? EU Enlargement and the Labour Market Outcomes of Immigrants in the UK

[7] Fassmann H (2009) European Migration: Historical Overview and Statistical Problems. Pp. 21-44 in Fassmann H, Reeger U, Sievers W: Statistics and Reality: Concepts and Measurement of Migration in Europe. Amsterdam University Press, Amsterdam

[8] Główny Urząd Statystyczny (2008) Informacja o badanich zasobów imigracyjnych w Polsce w 2008 r. Główny Urząd Statystyczny, Warszawa

[9] Ker D, Zumpe J, Blake A (2009) Estimating International Migration: An exploration of the definitional differences between the Labour Force Survey, Annual Population Survey, International Passenger Survey and Long-Term International Migration. Office for National Statistics

[10] Kępińska E (2008) Recent Trends in International Migration. The 2007 SOPEMI Report for Poland. CMR Working Papers No. 29/87, Warszawa

[11] Kępińska E (2007) Recent Trends in International Migration. The 2007 SOPEMI Report for Poland. CMR Working Papers No. 29/87, Warszawa

[12] Kupiszewska D and Nowok B (2008) Comparability of statistics on international migration flows in the European Union. In *International migration in Europe: Data, models and estimates*, Raymer J and Willekens F, eds., pp. 41-71. Chichester: Wiley.

[13] Kupiszewska D and Wiśniowski A (2009) Availability of statistical data on migration and migrant population and potential supplementary sources for data estimation. MIMOSA Deliverable 9.1 A Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

[14] Lemaitre G (2005) The Comparability of International Migration Statistics – Problems and Prospects, OECD Statistics Briefs, July 2005, No. 9. Organisation for Economic Co-operation and Development, Paris.

[15] Nowak L, Szałtys D, Kostrzewa Z, Pazderska A, Sobieszak A,Stańczak J, Jabłońska E (2007) Poprawa jakości i dostępności statystyki migracji zagranicznych. Raport finalny, Projekt Nr 12, Warszawa

[16] Nowok B (2010) *Harmonization by simulation: A contribution to comparable international migration statistics in Europe.* Amsterdam: Rozenberg Publishers.

[17] Nowok B, Kupiszewska D and Poulain M (2006) Statistics on international migration flows. In *THESIM: Towards Harmonised European Statistics on International Migration*, Poulain M, Perrin N and Singleton A, eds., pp. 203-231. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

[18] Nowok B and Willekens F (forthcoming) A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place* DOI: 10.1002/psp.624.

[19] OECD (2006). International Migration Outlook. Organisation for Economic Co-operation and Development, Paris.

[20] OECD (2008). International Migration Outlook. Organisation for Economic Co-operation and Development, Paris.

[21] Office for National Statistics (2009) Employment of Foreign Workers: Focus on Eurostat Ad Hoc Module 2008. Office for National Statistics.

[22] Official Journal of the European Union (2007) Commission Regulation (EC) No 102/2007 of 2 February 2007 adopting the specifications of the 2008 ad hoc module on the labour market situation of migrants and their immediate descendants, as

provided for by Council Regulation (EC) No 577/98 and amending Regulation (EC) No 430/2005.

[23] Official Journal of the European Union (2007) Commission Regulation (EC) Regulation (EC) No 862/2007 of the European Parliment and of the Council of 11 July 2007 on Community statistics on migration and international protection and repealing Council Regulation (EEC) No 311/76 on the compilation of statistics on foreign workers.

[24] Okólski M (1996) Recent Trends in International Migration. The 1995 SOPEMI Report for Poland. CMR Working Papers, Warszawa

[25] Okólski M (2009) Emigracja ostatnia? Scholar, Warsaw.

[26] Poulain M (1993) Confrontation des statistiques de migration intra-europèennes: vers une matrice complète? *European Journal of Population* 9(4):353-381.

[27] Poulain M (1999) International migration within Europe: Towards more complete and reliable data? Working Paper No. 37, Conference of European Statisticians, Statistical Office of the European Communities (Eurostat), Perugia, Italy.

[28] Poulain, M. and L. Dal (2008, June). Estimation of flows within the intra-eu migration matrix. Deliverable, GéDAP-UCL, Louvain-La-Neuve, Belgium.

[29] Poulain M, Perrin N and Singleton A, eds. (2006) *THESIM: Towards Harmonised European Statistics on International Migration.* Louvain: UCL Presses.

[30] Raymer J (2007) The estimation of international migration flows: A general technique focused on the origin-destination association structure. *Environment and Planning A* 12:371-388.

[31] Raymer J (2008) Obtaining an overall picture of population movement in the European Union. In *International migration in Europe: Data, models and estimates*, Raymer J and Willekens F, eds., pp. 209-234. Chichester: Wiley.

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

[33] Raymer J, de Beer J, 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(2):185-215.

[34] Ródenas C and M Martí (2007) Migration Estimation Based on the Labour Force Survey: An EU-15 Perspective. International Migration Review, Volume 41 No 1

[35] Robert CP, Casella G (2004) Monte Carlo Staistical Methods. Springer

[36] United Nations (1998) Recommendations on statistics of international migration. Statistical Papers Series M, No. 58, Rev.1, Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

[37] van der Erf R (2009) Typology of Data and Feasibility study. MIMOSA Deliverable 9.1 B Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

[38] van der Erf R (2010) Initial assessment of th quality of international migration data. IMEM Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

[39] Willekens F (1994) Monitoring international migration flows in Europe. Towards a statistical data base combining data from different sources. *European Journal of Population* 10(1):1-42.

[40] Willekens F (2008) Models of migration: observations and judgements. In *International migration in Europe: Data, models and estimates*, pp. 149-174. Wiley: Chichester.

[41] Woods RI, Rees PH (red.), Population structures and models, Allen and Unwin, London.