Independent Review of Methods for Distributing International Immigration Estimates to Regions

Jakub Bijak[1]

University of Southampton

Southampton, 4 January 2010

[1] School of Social Sciences, Southampton Statistical Sciences Research Institute (S3RI) and Centre for Population Change, University of Southampton, SO17 1BJ. Tel. 023 8059 7486, e-mail: J.Bijak@soton.ac.uk. The Review was carried out within the general framework for cooperation between the Office for National Statistics and S3RI, contract ONS-DMC4 (2009). All opinions expressed in this report are those of the author only and do not necessarily express the views of any of the institutions with which he is affiliated. I am very grateful to Jane Falkingham and James Raymer for their comments on the earlier draft, and to the staff of the ONS Centre for Demography for providing valuable insights into the methodology used by the ONS.

Executive Summary

This Review aims to examine and compare two methods for estimating the regional distribution of immigrants to England and Wales. The first method, currently used by the Office for National Statistics, is based on the International Passenger Survey, the Labour Force Survey and 2001 census data. Since its implementation in 2007, the method has undergone several modifications, and its further improvements, which are currently under development, include the use of administrative sources of data. An alternative method, developed at the University of Leeds, makes an explicit use of administrative data sources, with the aim to improve the estimates especially at the regional and local levels.

The comparison of the two methods reveals a trade-off between the consistency of definitions with respect to their compliance with the standard 12-months duration of stay criterion, and the coverage of all relevant populations. It seems that at the lower levels of spatial aggregation, such as local authorities, administrative sources provide much more up-to-date picture of immigration than estimates based on the 2001 census. The empirical differences between both methods at the higher levels, such as the Government Office Regions, require a separate enquiry, without which it does not seem possible to comment on the superiority of either of the two methods.

What is suggested instead is a compromise between the two approaches, taking into account both the needs of the user community, as well as realistic possibilities of providing objective and high-quality data by the ONS. Notwithstanding, this Review in principle supports the efforts to incorporate administrative sources of data into preparing regional and local migration estimates to the fullest extent possible, after ensuring that the underlying definitions are harmonised. However, if methodological changes are to be implemented by the ONS, this should be preceded by a due process involving consultations with the users of the data and, ideally, the academic community. It is also recommended that the future work concerning the use of administrative data sources for sub-national migration estimates should be preceded by detailed research in the following areas:

- Explanation of differences between the estimates yielded by various methods and based on different data sources, taking into account mechanisms of data collection.
- Development of methods for combining different, complementary data sources, in order to account for the uncertainty of measurement of international migration.
- Assessment of the possible bias of using the intentions-based residence and duration-of-stay declarations in the International Passenger Survey.

With respect to the latter, the Review raises concerns, which, although not in the direct focus of this enquiry, may require attention from the ONS and the user community. The most important issue regards the veracity of immigration 'totals' that are subject to redistribution. Since these numbers are based on intended rather than actual duration of stay, they may be biased, as it is likely the case with recent migration from the new EU Member States (the 'A8 countries'). While acknowledging that at present the data on intended stay provide the most comprehensive information on international migration, in the future they might be replaced or supplemented by more precise estimates, for example using the information from the 2011 census or the e-Borders scheme.

Since all challenges associated with the estimation are not trivial, the development of a robust methodology will critically depend on an adequate provision of resources. Given that, the completion of the recommended research tasks should be feasible within the framework of the Migration Statistics Improvement Programme, in the horizon of about four to five years. Meanwhile, a compromise on a feasible methodology is strongly recommended, bearing in mind that various sources of data are complementary, and that some measurement uncertainty is inevitable for such dynamic a process as migration.

1. Introduction

The aim of the current Review is to examine and compare two methods of estimating long-term immigration to England and Wales at the regional and local levels. The first method, currently applied by the Office for National Statistics (ONS, 2007a, 2008), uses the International Passenger Survey (IPS), the Labour Force Survey and 2001 census data to redistribute the total immigrant population, which should ideally depict, by definition, the number of immigrants who have entered the country in the past 12 months (ONS, 2009c). Here, this total is estimated from the IPS using the intended length of stay.

The alternative method, suggested by researchers from the University of Leeds (Boden and Rees 2010, forthcoming), proposes to enhance the quality of regional estimates by using administrative sources of data. These sources, assembled alongside several other ones in the form of the 'New Migrant Databank', include the National Insurance Number (NINo) data from the Department for Work and Pensions, the NHS data on the registrations of patients with General Practitioners (GPs), and the Higher Education Statistics Authority (HESA) data about students in the United Kingdom. This Review is chiefly based on the available published material, several unpublished documents, as well as, with respect to the most recent methodological developments, on personal communication with the ONS. It is worth stressing that, although the problems of disaggregation concern also emigration flows, their analysis remains outside of the remit of this study. For the same reason, short-term migration is not discussed either.

Apart from the current introduction, the Review is structured in four sections. The ONS method and the proposed alternative based on the New Migrant Databank are briefly summarised in Section 2. Subsequently, Section 3 provides an attempt to compare and evaluate both methods, by presenting their respective strengths and weaknesses from the theoretical and empirical points of view. Some key issues, which are not addressed by either method but seem to warrant further enquiry, are listed in Section 4. In particular, this concerns: (1) the reliability of the estimation of total immigration flows on the basis of the International Passenger Survey, and (2) the implicit presumption that various data sources used to redistribute immigrants need to be seen as alternatives, rather than as complementary. Finally, Section 5 summarises the main conclusions of the Review and puts forward some tentative recommendations for the future ONS work programme.

2. Methods for Regional Estimation of Immigration

This section is devoted to a brief description of both methods for distributing the international migration estimates among regions and local authorities of England and Wales. Subsection 2.1 discusses in brief the method currently used by the ONS, while Subsection 2.2 outlines the method based on the 'New Migrant Databank' developed at the University of Leeds. In both cases, the discussion focuses especially on the underlying definitions and coverage of various sources.

2.1. The ONS Method

The ONS method of estimating sub-national immigration flows by age and sex was developed firstly for the Government Office Regions (GORs) in England, as well as for Wales, then for the intermediate-level geographies (the so-called *New Migrant Geography for immigrants*, NMGi), and for local authorities (LAs). In principle, the method is based on the national total number of immigrants estimated from the International Passenger Survey. The IPS is a sample-based survey of people arriving to or departing from the UK through most of its airports or seaports, as well as through the Channel Tunnel. The problem with the IPS data, especially with respect to immigration, is that both the duration criterion (over 12 months of stay) and destination of migration are recorded on the basis of the intentions of respondents at the time of being surveyed at the port of entry, which may well differ from the actual duration and destination.

The method used by the ONS, implemented in 2007, partially addresses the issue of possible bias in reporting the intention-based destination of immigrants. To correct it, selected sub-national IPS distributions are replaced with the ones obtained from the 'rolling' average of the results of Labour Force Surveys (LFS) from five successive quarters (ONS, 2007a). The LFS is also a nation-wide survey, covering people aged 16 years or over who live in private households. The LFS excludes the populations of various communal establishments, among others, overseas students dwelling at the halls of residence.

It can be shown that at the GOR level, the spatial distribution yielded by the LFS roughly matches the one from the 2001 population census. At the same time this distribution is visibly different from the one obtained from the IPS, especially for London, where the difference is close to 8 percentage points (ONS, 2007a: 3). The differences between the IPS and the 2001 census for lower geographies, such as some local authorities in London (especially Kensington & Chelsea and Westminster) are also profound, but they can be related to the problem of distinguishing the factual (census) from the intended (IPS) places of residence.

The applied procedure estimation is as follows, with London treated separately from all other GORs (and Wales). Except for the Capital, the IPS estimates are used to calculate immigrant shares at the GOR and NMGi levels, using smoothed three-year averages, and the LA shares within each region are then taken from the 2001 census. For London, however, non-students and students (i.e. those who declared studies as the purpose for migration in the IPS) are treated separately. The IPS estimate for the latter is directly distributed across the LAs using 2001 census data for the student population. Subsequently, the non-students are redistributed into the NMGi regions using the LFS data, smoothed using three-year averages, and finally disaggregated into the LAs using the 2001 census. In all cases, both for London and other regions, the LA shares are subsequently calibrated to the regional totals at higher geographies (ONS, 2007a: 10–11).

Details of the calibration method are described in a separate working document (ONS, 2008). As noted by Boden and Rees (2010: 8), the use of the LFS for calibration is "the principal innovation of the [2007 improvements of the] ONS methodology". From the technical point of view, the methods seem sound, although several issues would require more clarification and updating, calling for a revised version of the documentation of the methodology. In such a revision, for example, the robustness of estimates for London and South-East of England (ONS, 2008: 4) could be assessed with respect to being a result of large population sizes.

At a more general level, the use of the LFS data to depict actual rather than intended place of residence of immigrants, has to be seen as a welcome improvement of the method of estimation at sub-national levels. However, from the documents under study it is not entirely clear why this method was applied to non-students only in London and not in other areas of England and Wales. The problem at this stage consists in merging two sources, which despite adhering to the 12-month criterion, rely on different definitions: one based on intentions (IPS) and one factual (LFS). This can bring about differences in the results. Moreover, another problem seems to be related to the coverage of the LFS data: the survey population excludes children below 16 years of age, who by definition are also not included among students. This problem can also slightly bias the outcomes.

A more serious issue relates to the use of 2001 census data for local authorities, which now can be seen as to a large extent outdated. Especially with respect to recent migration from the Central and Eastern Europe ('the A8 countries'), it was found that the spatial distribution of migrants was much different from the earlier settlement patterns of migrants from the same countries (e.g. Okólski and Grabowska-Lusińska, 2009). Besides, whilst the migrant network effects may be important, there are also many other factors

shaping the directions of migration, such as the availability of workplaces. In general, the links between migration flows and migrant stocks are far from simple and deterministic. To address this issue, the ONS is currently developing a model-based approach for LA-level estimation, using, among other, administrative data sources. This new methodology is envisaged for implementation by May 2010 (ONS, 2009c).

The documentation of the ONS (2007a) method concludes with suggestions for the future work, which, at the time of writing included the use of data on GP registrations (while correctly noting the problems with the three-month definition used therein), and performing a cluster analysis of 2001 census data for local authorities with respect to identifying communalities in the age and sex distributions. Especially the former has to be seen as potentially promising and worth addressing in a separate enquiry. It has to be noted that some of these issues are already being addressed within the framework of the Migration Statistics Improvement Programme, as for example in the model-based approach mentioned above (ONS, 2009c). Currently, plans for the future work of the ONS involve addressing some of the regional issues, developing methods for using micro-level administrative sources (see also Section 4), and improving the sex ratio estimation (*idem*).

2.2. The New Migrant Databank

The alternative method, proposed by Boden and Rees (2010) is based on the use of several administrative sources of data, with the aim to improve the regional estimates of immigration to England. A variety of such data has been assembled in the 'New Migrant Databank' at the University of Leeds (*idem*). These sources include, in particular, the National Insurance Number (NINo) data from the Department for Work and Pensions, NHS data on the registrations of patients with the General Practitioners (GPs), as well as the Higher Education Statistics Authority (HESA) data on overseas students.

The presented method also starts from the IPS estimate for the total inflow, or, more precisely, from an adjusted variant thereof, labelled by the ONS as the "Total International Migration" (TIM) estimate (for more details on TIM, see Table 1, Section 3). For sub-national divisions, two alternative models were proposed and compared. Model A makes exclusive use of the GP registration data at all spatial levels, including the GORs, intermediate geographies (NMGi) and local authorities. Model B is more complex, at least at the level of Government Office Regions, where three sources are used: the NINo data for workers, HESA for students and GP registration for all other immigrants. As pointed out by the ONS (2009c), although the HESA data in the presented version of the New Migrant Databank do not include the term-time postcode of the students, the refined version of the

data set, including this information, has recently become available and is used by the ONS. This can potentially allows for the use of the HESA data also at lower geographies.

The auxiliary variable used to distinguish between the three main groups (workers, students and other) is the IPS question on the declared aim of the visit. Thus, people who declare undertaking or looking for a job are assumed to be workers, those coming to pursue formal studies – as students, whereas all other immigrants are jointly treated as the 'remaining' category, which is subject to GP-based redistribution. Furthermore, the GP registration data are used as a universal source of information for the disaggregation at the NMGi and local authority levels (for a summary, see Boden and Rees, 2010: 25–26).

The method suggested by Boden and Rees (2010), and in particular Model B, provides a potentially very good coverage for those population groups, such as students, which could be difficult to target via nation-wide sample surveys[2]. The distributions at the lower geographical levels based on the GP registrations (a proxy for flows) rather than on the 2001 census stocks can be seen as more reliable, and are definitely more up to date.

On the other hand, it has to be noted that the administrative data on various population subgroups have been collected using different mechanisms, which were not primarily designed for statistical purposes. As a result, the definitions for some of the sources do not match the 12-months criterion: for HESA the duration-of-stay criterion may be six months or more in the case of postgraduate students (see ONS, 2009b, for details), while for the GP registrations the threshold equals three months, thus in reality depicting short-term migrants. The missing element of the analysis here is an evaluation of the impact of adopting different definitions and duration-of-stay criteria on the spatial distribution of immigrants. This issue is definitely worth addressing in a separate enquiry, in principle aimed at designing of a method of harmonisation of different sources by bringing them into a common standard of measurement. The first step could consist in comparing the spatial distributions of long-term and short-term migrants on the basis of the already available ONS estimates.

3. Comparison and Evaluation

In the current section, both methods described in Section 2 are compared and evaluated with respect to their respective strengths and weaknesses, both from the theoretical (Subsection 3.1) and empirical (Subsection 3.2) points of view.

[2] Although in the case of UK students, these are the parents who are often targeted in such surveys, this does not universally apply to overseas students.

3.1. Theoretical considerations

Both methods summarised in Section 2 make important improvements towards the IPS-based regional distribution of migrants in that they depart from the intention-based geographic attribution of migrants for at least some geographic levels. The main sources that are used (or may potentially be used) for the estimating the regional distributions by using either method are listed in Table 1.

Table 1. Comparison of coverage of various sources of data on immigration

Source	Population covered	Duration of stay	Remarks
Population census	In principle, the whole resident population of the UK (there are concerns about undercoverage, e.g. amongst migrant groups.)	12 months	Censuses are rare (the last one in 2001, next in 2011) and may not reflect interim changes in patterns, e.g. as seen for the A8 immigrants.
International Passenger Survey (IPS)	People entering and leaving the UK through a majority of air / sea ports (or the Channel Tunnel), but not from/to the Republic of Ireland	12 months (optionally: 3 months)	Sample survey ($p \approx 0.002$), in principle aimed to cover 90% of ports of entry or exit. Place of residence and duration of stay are based on intentions.
Total International Migration (TIM)	As in the IPS, but corrected for migration to/from Ireland, asylum seekers and 'switchers' of migrant status.	12 months (optionally: 3 months)	IPS, supplemented by the Irish data on migration to / from Ireland; Home Office data on asylum applications and estimates for 'switchers'
Labour Force Survey	Population aged 16 years or over, excluding residents of communal establishments (e.g student halls)	12 months	Sample survey, reasonable accuracy for higher levels of territorial division, but not below them.
NINo Allocations	Population aged 16 years or over, working or claiming benefits. Dependents are not registered	none	No duration of stay and no reporting of outflows or returns. Registration may be delayed since the actual immigration.
'Flag 4' GP registrations	Potentially all population groups: people making use of the NHS, who stay in the UK over 3 months	3 months	Definitional issues (short-term migration). Registration is not compulsory: healthy immigrants may not register. Registration may be delayed. Only the first registration is labelled as 'flag 4'.
Workers Registration Scheme (WRS)	A temporary register of working migrants from the A8 countries, except self-employed. Dependents are self-declared and may thus be double-counted	Different options (including 12 months)	Duration based on intentions (with many 'unknowns'). Charge of £90, which may contribute to undercoverage. Dependents joining later are not recorded. No outflows; records dropped after 12 months.
School Census (Department for Children, Schools and Families)	School-aged children in state schools, biannually.	none	Excludes independent schools. Collects information only on proxies (ethnicity and first language) rather than on migration.
HESA data on students	Overseas students, arriving for at least 12-month courses (or for 6-12 months in case of postgraduates)	6 or 12 months	Definitional difficulties. Possible overlap with workers and patients. Excludes students at non-HESA institutions (e.g. language schools)
Home Office: work permits, immigration control	Non-EEA immigrants only	none	No outflows. The immigration control data are available at the national level only (Boden and Rees, 2010)

Source: Adapted from ONS (2007b, 2009a,b), IPPR (2008) and Boden and Rees (2010); own elaboration

It has to be noted that in the case of both methods the respective estimates are benchmarked to the IPS totals. Both also follow the geographical hierarchy, from the GORs through NMGi to local authorities. Their distinct features, both strengths and weaknesses, stemming from the use of different sources of data, are summarised in Table 2.

Table 2. Strengths and weaknesses of two estimation methods of sub-national immigration

Method	The ONS Method	The New Migrant Databank
Strengths	• LFS: *de facto* definition, consistent with the 12-months UN criterion • The same mechanism of data collection for most subgroups	• Good coverage for various groups of population (students, workers) • More up to date distribution at the lowest geographies, more adequate for specific groups (e.g. migrants from A8 countries)
Weaknesses	• LFS excludes foreign populations from communal establishments, e.g. student halls of residence • LFS sample size too small for estimates at lower geographies • The use of 2001 census data for distributions at the LA level is most likely outdated, e.g. with respect to A8 migrants • IPS definition based on intentions for the total, GORs and NMGi (the same benchmark for both methods)	• The definitions for some sources (HESA, GP registrations) do not match the 12-months criterion and the data might thus require additional calibration • Data on various population subgroups collected using different mechanisms, not designed for statistical purposes • No outflows recorded in some sources (especially NINo) • IPS definition for the total based on intentions on the duration of stay (the same benchmark for both methods)

Source: Own elaboration based on ONS (2007a, 2008, 2009b) and Boden and Rees (2010)

The main conclusion from the comparison of methods seems to be a trade-off between the consistency of definitions with the standard UN 12-months criterion, applied by the ONS, and a better coverage of all or most relevant populations, manifested in the method developed at the University of Leeds. The strengths of the current ONS method lie in the use of the same time criterion consistently for all groups of population, with an attempt to correct for the intentions reflected in the IPS data by using LFS-based calibration. On the contrary, the strengths of the New Migrant Databank methodology consist of a coverage of specific populations excluded from the Labour Force Survey (such as halls-based foreign students), and of the methodological features of estimation at the lowest levels of spatial disaggregation. Especially for the local authorities, information on flows based on the GP registrations, exact to the definition applied, seems more up-to-date than 2001 census data on population stocks, used as a proxy for the allocation of flows.

This trade-off has implications for the usefulness of particular methods for the users of data. From the purely theoretical point of view, it seems that for the local authorities, the information obtained from administrative sources (via the New Migrant Databank) is likely more complete. The concern here, however, is the lack of recording of outflows in some of the sources (NINo), which may lead to an overestimation of the number of actual immigrants. At higher geographies, consistent definitions of the IPS and LFS, at least with respect to the time criterion, seem to be an advantage of the ONS method. However, as shown by Boden and Rees (2010), and discussed further in Section 3.2, the two methods yield different results even at the level of the Government Office Regions and constituent countries of the UK. Before making further inferences, these differences need to be explained by referring to the characteristic features of specific subpopulations, as it was done for example for the students in London in the ONS (2009b) study.

Notwithstanding, in the longer run more work on the use of administrative sources seems indispensable, subject to more thorough feasibility checks. Such course of action has already been stipulated in several earlier reports (e.g. ONS, 2007b; UK Statistics Authority, 2009). To overcome some of the theoretical shortcomings of administrative sources, an appropriate mechanism for harmonising definitions across sources should be developed. Some suggestions are provided in Section 4.2, as well as in Raymer *et al.* (2007) and Raymer and Abel (2008). In any case, efforts should be made to incorporate some of the important administrative sources of data into regional and local migration estimates. Noteworthy, this task is included in the upcoming changes in the ONS methodology envisaged for 2010 (ONS, 2009c).

3.2. Empirical results

Further to the theoretical differences underpinning the two methods, their authors found them dissimilar also with respect to the actual values of the estimates. On the GOR level, the New Migrant Databank estimates from Model B were higher than their respective counterparts computed using the ONS method especially for West Midlands, by 33%, and London, by 12% (Boden and Rees, 2010: 27). The opposite held especially for the East of England, the ONS estimate being higher by 23%, for Yorkshire and the Humber (by 21%) and the South West (by 19%). In the remaining four cases, the difference did not exceed 10% (*idem*). Within GORs, the differences at the level of NMGi and local authorities varied and could reach extreme levels, as for example for East Staffordshire (West Midlands), where the Model B estimate was by 252% higher than its ONS counterpart (*idem*: 28–29).

Based on the purpose of stay declared in the IPS, over a quarter of migrants to the United Kingdom are students (Boden and Rees, 2010: 11). Given the size of this group, an attempt to shed light at some of the differences between data sources on student flows was undertaken in the recent material prepared by the ONS (2009b). To start with, the authors of the document acknowledged a definitional weakness of the LFS with respect to the student population. To correct for it, an amended approach was proposed, whereby students would be redistributed into the regions of England and Wales according to the HESA data, while all other population groups would follow a LFS-based pattern of spatial distribution. Interestingly, an application of the amended method led to a further decline of the allocation of migrants to London, contrary to the findings of Boden and Rees (2010) obtained on the basis of the New Migrant Databank statistics (cf. ONS, 2009b: 12).

For London, the ONS (2009b) report offers some tentative explanations for the discrepancies in estimates, such as higher propensity to live in private accommodation amongst London students. Secondly, it was noted that the 'student' category as recorded by the IPS may be broader than its HESA counterpart. In general, however, it seems that without more detailed local knowledge of the regional specifics, it may be difficult to arrive at any definite conclusions as to the superiority of any particular method or data source. Boden and Rees (2010: 13) suggested that in order to explain the exceptionally large differences between the two methods, local consultations may be necessary.

Besides, both Boden and Rees (2010), and the ONS (2009b: 12) report concur that the use of administrative data sources can potentially improve the quality of sub-national estimates of international immigration. Both also suggest that some additional research pathways should be explored before such use can be directly utilised in preparing the official population estimates for the UK. Boden and Rees (2010: 14) propose to extend the New Migrant Databank by additionally including the School Census data. Secondly, they recommend exploring the possibilities offered by individual-level data matching[3]. The extension of research interest to emigration, as well as short-term migration flows (between 3 and 12 months) was also suggested (*idem*). The ONS (2009b: 12) report, in turn, calls for a "more thorough investigation" of the potential of administrative sources, and notes that the Office "intend[s] to make this a priority in the next phase of the Migration Statistics Improvement Programme" (*idem*). Since the two proposals are not mutually exclusive, it can be argued that both directions of research could be beneficial for improving the quality of the estimates. Some additional ideas are suggested in Section 4.

[3] For a discussion of some ideas concerning the possibilities offered by matching, see Section 4.3.

4. Further Work

The current section is devoted to two issues that have not been fully addressed by either of the methods under study, but may be worth addressing in separate research enquiries. Firstly, the reliability of estimation of total immigration flows on the basis of the International Passenger Survey (IPS), is discussed in Subsection 4.1. Subsequently, a more fundamental issue is addressed in Subsection 4.2, about whether different sources of data should be treated as conflicting, or rather as complementary pieces of statistical information on immigration to England and Wales. Finally, some additional suggestions on possible methods for combining of sources are provided in Subsection 4.3.

4.1. Totals

An important, general issue with respect to the estimates of immigration seems to concern not the spatial redistribution of the aggregate IPS totals, but rather these totals *per se*. Both methods under review aptly address the problem of intentionality of IPS answers with respect to the place of residence. However, a similar problem arises with respect to the intended duration of stay, which has a profound impact on the totals. As noted by Kupiszewski (2002), especially in the context of such a dynamic process as migration, intentions usually are a very poor predictor of the actual duration of residence.

One recent example of differences between the IPS and the administrative sources concerns migration from Central and Eastern Europe (the 'A8 countries') following the EU enlargement in 2004. According to some insights available from qualitative studies, many among the A8 migrants adopt a strategy of 'intentional unpredictability' with respect to their duration of stay in the UK (Eade *et al.*, 2006). As noted by the IPPR (2008: 17–18), between 1 May 2004 and 31 December 2007, there were 796 thousand applications of A8 citizens to the Workers Registration Scheme (WRS). According to the ONS data, the IPS total inflow from the A8 countries in the same period, even after corrections producing the amended Total International Migration (TIM) estimate, thus already accounting for 'switchers' from visitors to migrants and vice versa, was only 333 thousand people.

A similar finding has been also reported by Boden and Rees (2010: 22), who showed the increasing gap between the TIM and other data sources from the New Migrant Databank after 2004. This gap can be to some extent attributed to the presence of short-term migrants and the related issue of not deregistering them from several administrative sources upon departure. However, given all the difficulties involved, the result is that research studies of the volume of post-accession migration from the A8 countries to the

United Kingdom are scarcely based on the IPS estimates alone, but rather a combination of various administrative and other sources, such as the IPS, LFS, WRS, NINo and other[4].

A possible way to overcome the above-mentioned difficulties would be to depart from the IPS as the sole basis for the estimation of annual international migration flows. A solution, mentioned in the review for the UK Statistics Authority (2009), would be to use the data on the actual rather than intended duration of stay, which might become available from the e-Borders scheme, initially scheduled to be operational two to three years after the 2011 population census (*idem*)[5]. Before the e-Borders data become fully operational, the IPS can be supplemented, if needed for example for specific research purposes, by the administrative sources, as it indeed is a common practice already today. Prior to that, however, the IPS data, despite the shortcomings mentioned above, are likely to remain the most comprehensive source of data on international migration available for the UK. Besides, the use of IPS as a benchmark currently allows for harmonising the sub-national estimates of flows, at least in terms of their overall scale. Hence, the remarks made in this section refer predominantly to the future work and changes that anyway seem inevitable, rather than suggesting an additional element to the current Migration Statistics Improvement Programme (cf. UK Statistics Authority, 2009).

4.2. Conflicting sources or complementary sources?

From the comparison of methods and their empirical results, outlined in Section 3, it becomes apparent that there are visible differences between regional estimates obtained by using different methods. These differences depend on the data sources used, and can be quite substantial in size, especially for some regions. The disparities can be attributed either to different definitions underlying particular sources, which lead to variation in coverage, or to methodological considerations, such as small sample sizes in surveys.

Given that various sources of data used in both methods essentially measure slightly different processes, the differences are not surprising. However, for the same reason, since all these sources have been originally designed for other purposes than the measurement of migration, the very quest for an 'optimal method' may become problematic. An alternative point of view, which can be adopted to tackle this problem, would be to treat

---

[4] Next to the IPPR (2008) report, the relevant examples concerning Polish migration include the estimates of the Central Statistical Office of Poland (CSO, 2008), as well as the ones included in the research published by Okólski and Grabowska-Lusińska (2009).

[5] Given the recent questions about the legality of the e-Borders scheme with respect to intra-EU movements of people (House of Commons Home Affairs Committee, 2009: § 44-50), it might be possible that the implementation of the system is either delayed, or limited in scope to non-EU movements.

alternative sources as providing complementary rather than conflicting evidence on the actual distribution of immigrants. By adopting such a view, the focus of the enquiry would shift from the provision of a single statistic (point estimate) of migration to the adequate description of the uncertainty of its measurement.

In migration research, attempts to correctly describe uncertainty for a system of flows, on the basis of various sources of data, were undertaken only very recently [6]. The prototype studies on combining macro-level data from different countries, however, were already carried out before, as discussed for example in the paper of Brierley et al. (2007). Some mechanistic attempts to harmonise different data sources on migration were also undertaken within the framework of the Eurostat-funded project MIMOSA (*Migration Modelling for Statistical Analyses*, for more details, see e.g. Raymer and Abel, 2008).

The main advantage of such approaches consists in an explicit acknowledgement of imperfect measurement as yet another source of uncertainty, next to, for example, the sampling error in surveys (IPS or LFS). The Bayesian statistical approach, suggested in this context by Brierley et al. (2007), which produces the probability distributions of quantities under study (here: region-specific shares of immigrants), is a natural and coherent method of analysis. An additional advantage is that such methods do not require individual-level data, being based on macro-level aggregates instead. The prerequisite for their practical application, however, is an agreement about their usefulness among the data users. Since statistical methods of combining macro-level data are only currently being developed, some alternative ideas of more straightforward research strategies are suggested below.

4.3. Combining sources: Proposal for a research agenda

As to other research pathways that can be explored with the aim of combining various data sources, both macro-level and individual-level approaches can be mentioned. Within the former group, perhaps the most straightforward, but also the most crude method would consist in the averaging of results (shares) yielded by different data sources. In such method, the weighted averages of source-specific shares are computed for each region under study. Subsequently, the averaged shares can be adjusted (for example proportionally) in order to ensure the summation of regional counts to the overall total.

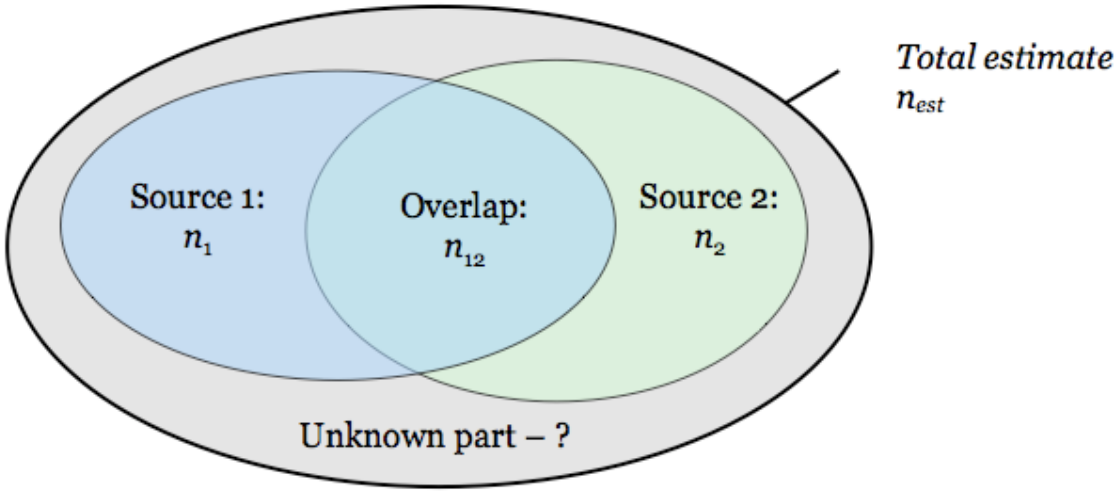
---

[6] The S3RI, University of Southampton jointly with two international partners (Netherlands Interdisciplinary Demographic Institute and the University of Oslo) have just started a two-year research project IMEM (*Integrated Modelling of European Migration*), funded by the NORFACE network. The aim of the enquiry is to harmonise and model migration flows among the EU and EFTA countries applying the Bayesian statistical aproach. One of the challenges of this project consists in combining information from different sources using formal statistical methods within a framework of a common, hierarchical model.

The weights given to different sources are by necessity to some extent judgemental and reflect beliefs in the accuracy of particular sources. In practical applications, they can be obtained for example from a panel of experts in the field. Under no preference for any particular source, equal weighting can be applied. In the context of migration, averaging was previously used for example to distinguish between different drivers of population flows (Harker, 1986). This method was also applied to the preparation of official scenarios of international migration for the EU-15 countries, within the EUROPOP-2003 population projections of the Eurostat (Lanzieri, 2004).

Alternatively, should individual-level data from different sources (e.g. from the LFS and NINo) become available with common identifiers, such as name and address, another possible method consists in the micro-matching of data between different sources and analysing their overlap. From the statistical point of view, such an approach would enable the application of 'dual system analysis' also known as 'capture-recapture method' (Marks, Seltzer and Krótki, 1974, after: Alho and Spencer, 2005: 26). In its simplest, two-source form, the estimate of the total, n_{est}, can be obtained using the estimates from both sources, n_1 and n_2, and their overlap, n_{12}, as: $n_{est} = n_1 n_2 / n_{12}$, with variance $\mathrm{var}(n_{est}) = n_1 n_2 (n_1 - n_{12}) \cdot (n_2 - n_{12}) / (n_{12})^3$ (*idem*). This formula requires an assumption of independence of both sources, however there also exist extensions to situations with (partial) dependence, more than two sources, etc. A simple scheme of partitioning of the unknown total into known source-specific subtotals, their overlap, and the unknown remainder, is shown in Figure 1.

Figure 1. Schematic framework for a two-source dual system estimation



Source: Own elaboration

It is worth noting that the dual system estimation was already used by the ONS to assess the coverage of the 2001 population census on the basis of the post-enumeration

survey (for methodological details, see e.g. Brown et al., 2000). Also the review prepared for the UK Statistics Authority (2009: 92–93), citing Mayhew and Harper (2009), suggests that even without applying statistical methods, individual-level linking of sources is a promising method for producing reliable estimates at various levels of spatial aggregation.

Moreover, it is also possible to estimate the regional shares on the basis of multiple sources even without linking individual-level records. A relevant method could for example consist in assessing the overlap between various populations covered by different data sources (such as those listed in Table 1), on the basis of general surveys or similar sources of information. An example can be the assessment of the number of students living in private households, which could then be used for merging the HESA and LFS datasets. Another option would be to try to evaluate the bias of the GP registration data, on the basis of, for instance, the General Household Survey or the British Household Panel Survey.

Regardless of the method used for estimating the overlap, the information from individual sources could subsequently form as a basis for conducting a dual (or indeed multiple) system estimation, as outlined above. Alternatively, this information can be used in a deterministic fashion, by applying a simple balance equation, e.g. $n_{est} = n_1 + n_2 - n_{12}$. The use of balance equations tacitly assumes that the sources under study exhaust the coverage of the whole population in question, with no unknown part remaining (see Figure 1). This approach effectively leads to estimating the "minimum confirmed population", suggested as an analytical framework by Mayhew and Harper (2009: 9) in the context of linked multiple sources of data.[7]

5. Conclusions and Recommendations

This Review confirms that both methods under study constitute important steps towards improving the quality of sub-national estimates of international migration flows. An analysis of their strengths and weaknesses identified a trade-off between the adherence to common definitions with respect to the duration-of-stay criteria but at a risk of lower coverage of migrant populations (in the ONS method), and more complete and up-to-date coverage, at the expense of some definitional differences (in the New Migrant Databank). This trade-off is visible especially at higher levels of spatial disaggregation (GORs and NGMi), where without more in-depth studies in the future it would be very difficult to make a definite statement on the superiority of any of the two approaches.

[7] As an example, Bijak et al. (2007) corrected the estimate of the population size of Warsaw, using, among others, administrative data on students, business reporting data on the number of persons employed in the city, as well as survey data on the share of students who work, which enabled to estimate the overlap.

At the level of local authorities, however, the preferred solutions is the one based on administrative sources, involving for example such methods as those proposed by Boden and Rees (2010) or the model-based approach currently being under development in the ONS (2009c). Due to the relatively long time span since the 2001 census, interim events that changed the geography of migration (such as the 2004 EU enlargement), as well as the ambiguities of the impact of migrant stocks on subsequent flows, the use of 2001 census data does not seem fit for purpose.

At the current stage of developments, given the limitations of various data sources, a compromise between both approaches is therefore required. Under such a compromise, ideally both the ONS as a data providing authority and the user community should be satisfied that the methodological state-of-the-art is of highest possible quality, especially given the constraints on the resources available for data provision. Hence, any changes to be implemented by the ONS should be preceded by a due process, involving consultations with the users of the data and, ideally, with the academic community.

For the future work on the methodology for distributing international migration estimates among the UK regions, the following tentative recommendations can be put forward. As suggested before (ONS, 2009b; UK Statistics Authority, 2009; Boden and Rees, 2010), the use of administrative data sources for sub-national migration estimates should become an important part of the Migration Statistics Improvement Programme. However, a practical application of new methods based on combined sources (and ideally also adjusted totals) to the calculation of official migration and population statistics for the UK should be ideally preceded by thorough research in the following areas:

- Studies should be conducted in order to shed more light on the differences between the estimates yielded by various methods and based on different data sources. Better understanding of these differences can contribute to a more robust methodology, which could, at least partially, account for the differences in definitions and coverage. A good example is the study of HESA students, with focus on London (ONS, 2009b).
- A useful starting point could be a comparison of regional distributions yielded by various administrative sources in 2010-2011 with the results of the 2011 population census. Such a comparison can be especially informative at the local authority level.

- Various data sources, administrative and survey-based alike, should be treated as complementary, although imperfect, pieces of information on international migration. Methods for combining them, in order to account for the uncertainty of measurement, should be developed. Tentative suggestions are presented in Section 4.
- The possible bias of using the intentions-based residence and duration-of-stay declarations in the International Passenger Survey for the purpose of estimating total migration flows to and from the UK should be assessed. This can be done for example by using the 2011 population census or e-Borders data, if they become available.

Since all challenges associated with the estimation are not trivial, the development of a robust methodology for spatial redistribution of international migrants in the UK will critically depend on an adequate provision of resources. Given that, the completion of the related research tasks should be feasible within the framework of the Migration Statistics Improvement Programme, in the horizon of about four to five years. Meanwhile, a compromise on a feasible methodology, as suggested above, is strongly recommended.

References

Alho J and Spencer B (2005) *Statistical Demography and Forecasting*. Berlin-Heildelberg: Springer.

Bijak J, Kicinger A and Kupiszewski M [co-operation: P Śleszyński] (2007) *Studium metodologiczne oszacowania rzeczywistej liczby ludności Warszawy* [*Feasibility Study for the Estimation of the Actual Population Size of the City of Warsaw*]. Working Paper 2/2007. Warsaw: CEFMR. Available at http://www.cefmr.pan.pl/docs/cefmr_wp_2007-02.pdf (Access: 7 November 2009).

Boden P and Rees P (2010) Using administrative data to improve the estimation of immigration to local areas in England. Forthcoming in *Journal of the Royal Statistical Society A*. Available at http://arxiv.org/pdf/0903.0507 (Access: 20 October 2009).

Brierley M J, Forster J J, McDonald J W and Smith P W F (2007) Bayesian estimation of migration flows. Pp. 149–174 in J Raymer and F Willekens (eds.) *International Migration in Europe: Data, Models and Estimates*. Chichester: John Wiley.

Brown J, Chambers R and Cruddas M (2000) *Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey*. Report for the National Statistics One Number Census Steering Committee. University of Southampton. http://www.statistics.gov.uk/Census2001/pdfs/sc0003a.pdf (Access: 7 November 2009)

CSO (2008) *Informacja o rozmiarach i kierunkach emigracji z Polski w latach 2004–2008* [*Information about the volume and directions of emigration from Poland, 2004–2008*]. Warsaw: Central Statistical Office. http://www.stat.gov.pl/cps/rde/xbcr/gusPUBL_lud_infor_o_rozm_i_kierunk_emigra_z_polski_w_latach_2004_2008.pdf (Access: 7 November 2009).

Eade J, Drinkwater S and Garapich M P (2006). *Class and Ethnicity: Polish Migrant Workers in London*. ESRC Project Report. CRONEM, University of Surrey, Guildford.

Harker P T (1986). The use of expert judgments in predicting interregional migration patterns: An analytic hierarchy approach. *Geographical Analysis*, 18(1): 62–80.

House of Commons Home Affairs Committee (2009) *The E–Borders Programme*. Third Report of Session 2009–10. London: The Stationery Office. http://www.publications.parliament.uk/pa/cm200910/cmselect/cmhaff/170/170.pdf (Access: 22 December 2009).

IPPR (2008) *Floodgates or Turnstiles? Post-EU enlargement migration flows to (and from) the UK*. Institute for Public Policy Research. Available at http://www.ippr.org.uk/publicationsandreports/publication.asp?id=603 (Access: 3 October 2009).

Kupiszewski M (2002) How trustworthy are forecasts of international migration between Poland and the European Union? *Journal of Ethnic and Migration Studies*, 28(4): 627–645.

Lanzieri G (2004) EUROPOP 2003: Methodology for Drafting International Migration Assumptions in the EU-15 Member States. Unpublished paper for the meeting of the Working Group on Population Projections. Eurostat, Luxembourg, 15–16 July.

Marks E S, Seltzer W and Krótki K (1974) *Population growth estimation: A handbook of vital statistics measurement*. New York: Population Council.

Mayhew L and Harper G (2009) Making Use of Local Administrative Data For Population Estimates and Service Planning. Paper presented at UPTAP Workshop, Leeds 23–25 March. Slides: http://www.uptap.net/wordpress/wp-content/uploads/2009/05/les-mayhew-uptap-presentation.pdf (Access : 8 November 2009).

Okólski M and Grabowska-Lusińska I (2009) *Emigracja ostatnia?* [*The last emigration?*] Warsaw: Scholar.

ONS (2007a) *Improved Methods for Estimating International Migration Geographical Distribution of Estimates of In-migration*. London: Office for National Statistics. http://www.statistics.gov.uk/downloads/theme_population/Geog_distn_in-migs.pdf (Access: 16 October 2009).

ONS (2007b) *A Review of the Potential Use of Administrative Sources in the Estimation of Population Statistics* (Amended). London: ONS. http://www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/admin.pdf (Access: 6 November 2009).

ONS (2008) *Improved Methods for Estimating International Migration: The Use of Calibration in Estimating International In-migration to UK Countries and Regions of England*. Titchfield: ONS. http://www.statistics.gov.uk/downloads/theme_population/Callibration.pdf (Access: 16 October 2009).

ONS (2009a) *Local Area Migration Indicator Metadata*. Titchfield: ONS. http://www.statistics.gov. uk/downloads/theme_population/metadata.pdf (Access: 12 October 2009).

ONS (2009b) Potential Use of HESA data to improve the regional distribution of international migrant inflows. Unpublished document. Titchfield: ONS.

ONS (2009c) Personal communication, 19 November and 17 December 2009.

Raymer J, Abel G (2008) The MIMOSA model for estimating international migration flows in the European Union. Paper for the Joint UNECE/Eurostat Work Session on Migration Statistics, Geneva, 3–5 March. http://www.unece.org/stats/documents/ece/ces/ge.10/2008/wp.8.e.pdf (Access: 7 November 2009).

Raymer J, Abel G 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* A, 170: 891–908.

UK Statistics Authority (2009) *Migration Statistics: the Way Ahead? Report by the UK Statistics Authority*. Report 4. London: UK Statistics Authority.

■