Internal Migration Methodology: An Improved Method of Estimating Student Migration

1. Summary

This note describes a proposed improvement to the methods used in estimating internal migration within England and Wales. Comments on the proposed approach are welcome and should be sent to migstatsunit@ons.gov.uk.

2. Background

The Local Authority District (LAD) internal migration estimates produced by ONS are primarily based on Patient Register data. Comparing individuals' addresses on this year's and the previous year's Register allows us to identify 'transitions' over the year. These are scaled up, using other data extracted from NHS systems, to allow for moves where the person is not living at their migration destination at the end of the period (for example, where the person has died between the date of migration and the end of the period).

It had long been recognised that a weakness of this method is that it did not adequately capture migration of students. This is because students are less likely than most other population groups to immediately re-register with a GP after moving (to or from their place of study).

This problem was addressed as part of the Improving Migration and Population Statistics Programme. This resulted in an improved methodology which uses Higher Education Statistics Agency data to estimate student migration flows[1]. That new method was first used in 2010, and the current series of internal migration estimates from 2002-2011 has now been calculated on this basis (these are the estimates used in the current 'rolled forward' population estimates based on the 2001 Census).

Though this work was a significant step forward it was acknowledged that the method – which relies on estimates of migration derived separately from Patient Register data and HESA data –

[1] http://www.ons.gov.uk/ons/guide-method/method-quality/imps/msi-programme/communication/improvements-mid-2008/methodology-papers/student-adjustment-detailed-methodology.pdf

could be further improved by linking the two datasets to identify the most reliable address[2] for each individual and to derive the migration estimates from that combined dataset. This approach could lead to more accurate estimates and to a simpler methodology – which would provide greater transparency to users.

3. Method

3.1 Preparing Data

The HESA data is filtered before being matched to the Patient Register data. This filtering removes:

- records with domicile not in England and Wales;
- dormant records (that is, records not related to a current student)
- records with Output Area code of domicile missing
- records containing duplicate values of matching variables (that is, date of birth, sex and domicile OA code).

3.2 Linking Datasets and Identifying Area of Usual Residence

The Patient Register and HESA datasets are linked on date of birth, sex and Output Area of residence/domicile (this is the 'home' (rather than term-time) Output Area on the HESA data).
The linked dataset now contains two Output Area codes for matched records – the Patient Register code (which must necessarily be the same as the HESA 'home' code if the record has been matched) and the HESA term-time code[3]. It also contains the date on which the patient registered with a GP. We adopt the HESA term-time code as the 'correct' code unless the GP registration date is later than the start of the academic year, in which case we use the Patient Register code.

3.3 Adjusting for Changes in Address at the End of Studies

Whilst the HESA data is very useful in identifying moves at the start of studies, it does not contain reliable information on moves at the end of the studies. This means that a student who does not update their Patient Register record will be recorded, using the above methodology, at their term-time address while they appear on the HESA dataset and then immediately revert to their pre-study 'home' address when they finish their studies and disappear from the HESA data. As a proportion of students will actually stay in their study area after finishing their studies, this immediate reversion would have the tendency to overestimate outflows of 22/23 year olds from student areas and underestimate outflows in the mid/late twenties.

[2] The method is actually based on Output Area code rather than full address. To improve the readability of this paper, 'address' has been used in some places where 'Output Area code' is strictly correct.
[3] As term-time address was not provided on the HESA data prior to 2007/08, this will be imputed on the basis of Campus ID when developing a back series of estimates from 2002/03.

To avoid this, 'moving out factors' are calculated and applied to the data. The idea behind these is as follows: if someone was at a HESA term-time address last year but is not on the HESA dataset this year then they will revert to their Patient Register address with a probability specific to their sex and LAD of study. These probabilities are calculated by looking at tendency to move for students who have updated their Patient Register record.

Example:

John Jones registered with a GP in Gosport in 2001. He went to University in Cambridge in September 2006 and finished his studies in June 2009. His Patient Register record has not been updated since he first registered.

Up to mid-2006 John will have been assumed[4] to live in Gosport. When he started University he will have appeared on the HESA dataset applying from September 2006, and, since he had not registered with a GP since then, he was assumed to live at the HESA term-time address at mid-2007; and similarly for 2008 and 2009 (as a point of practicality we assume that students are usually resident at their term-time address at the end of June of their final year of study unless we have specific evidence (from the Patient Register) that this is not the case).

In September 2009 he did not appear on the HESA dataset. We have estimated separately that male students who studied in Cambridge have (say) a 70% probability of moving out of that LAD each year. By generating a random value and comparing with this probability we decide that though John does not appear on the HESA dataset we will retain his study address as his 'correct' address for mid-2010.

'Given that' John did not move out of Cambridge in the year to mid-2010, there is a 70% probability that he will move out in the year to mid-2011. Again we test against a random number and, on this occasion, replace the HESA address with the Patient Register address in Gosport. John's address will now change only if he changes his address on the Patient Register record or if he becomes a student again.

[4] That is, assumed for the purposes of the internal migration estimates.

3.4 Calculating Probabilities of Moving Out at the End of Studies

As noted above, the probability that an ex-student of a particular sex and LAD of study, who has not had their Patient Register record updated, moves out of that LAD during a year is calculated by looking at the corresponding group who have had their Patient Register record updated. In practice, this probability is generally somewhat higher in the first year rather than the second year. In order to derive a single probability that can be applied to lags of any length we have taken the square root of the product of the probability of leaving with lags of 1 and 2 years (thus, the probability of leaving after two years is unchanged). The impact of this methodological inaccuracy (moving slightly too few people out in the first year but then compensating in the second year) is likely to be small compared to other uncertainties in estimating these probabilities.

It is assumed that, unless we have actual evidence from the Patient Register of an updated address, the destination LAD of an ex-student moving out is their previous LAD recorded on the Patient Register. It would be possible to replace this assumption with some statistical model of destination LAD given origin LAD (as used in the current method). This might be expected to, very slightly, improve the accuracy of the estimates, but would also increase the complexity of the methods and uncertainty surrounding the estimates. As it would not significantly address the issue of most concern – namely that ex-students are moved out of their area of study at an appropriate rate – we do not propose to introduce the additional complexity of modelling destination LAD.

3.5 Identifying Migrants and Deriving the Final Migration Estimates

Identifying migrants is carried out in the same way as the existing method, but using the dataset described above (incorporating HESA term-time addresses where these are thought more reliable than the Patient Register addresses). For 2011, for example, the 2011 dataset is matched to the 2010 dataset, and records where the Local Authority of address has changed are identified as migrants.

Once these flows of identified migrants have been derived, these are scaled up to allow for moves where the person is not living at their migration destination at the end of the period, as described in the ***Background*** section. These scaling factors (by age, sex and origin and destination area) are calculated by comparing NHS Central Register (NHSCR) reports on all moves within the year to Patient Register data on moves where the person was present at both the start and the end of the year. In this situation it would be incorrect to use the improved dataset, as the additional identified student migrants will not appear in the moves within the year data, and the scaling factors would thus be incorrectly low.

3. Advantages of Proposed Approach

Accuracy: Though we do not yet have quantitative evidence for this, we can reasonably expect the estimates produced using this method to be more accurate than those produced using the existing method.

Simplicity: The method is much simpler than the current method and does not require separate adjustments to avoid double-counting or for 'within study moves'. Amongst other advantages, this simplicity should make the methodology more 'transparent' and the estimates easier to explain.

Robustness: Whilst inaccuracies may remain where individuals have not updated their Patient Register records, these inaccuracies will (almost exactly[5]) correct themselves over time. In the example above, if John Jones had moved to Portsmouth rather than Gosport after completing his studies, we will have incorrectly 'counted' him as a move from Cambridge to Gosport in the year to mid-2011. When John finally updates his Patient Register record, however, he will be counted as a move from Gosport to Portsmouth – so, the final net effect is correct even though the intermediate move was incorrect.

3. Next Steps

This proposed methodology is being provisionally implemented as part of the internal migration redevelopment project due to be completed at the end of 2012. Amendments to the draft methodology will be considered as a result of any comments from users. The development will be accompanied by the collection of quantitative evidence on whether the method is robust and whether assumptions are justified. It is planned that the improved method is used in the year to mid-2012 estimates to be published in mid-2013, with the likelihood that a revised back series for 2002-2011 is also published.

We will also proceed with the assessment of the PDS data source, with a broad expectation that this will replace the Patient Register and NHSCR data sources for estimates published from mid-2014 onwards. We would provisionally expect that estimates based on PDS data would also include an adjustment to allow for student migration not captured on NHS data sources.

Internal Migration Team
October 2012
Contact: migstatsunit@ons.gov.uk

[5] Very small discrepancies could occur due to differences in NHSCR scaling factors for the different moves.