Measuring uncertainty in the Population Estimates for Local Government Areas in England and Wales

Fulton, Ruth
Office for National Statistics
Segensworth Road,
Titchfield, PO15 5RR, UK
E-mail: ruth.fulton@ons.gov.uk

Overview of population estimates methodology

In England and Wales population estimates for local government areas (local authorities[1]) are calculated using the cohort component method. Basically, the previous year's population is aged-on by one year and then adjusted for births, deaths, internal and international migration[2]. International migration estimates are based on data from the International Passenger Survey (IPS), data on asylum seekers and their dependents, and estimates of migrant and visitor switchers[3]. In addition, adjustments are made for special population groups that are not captured by the internal or international migration estimates: members of the armed forces, prisoners and pupils in boarding schools. The estimation process is repeated each year, starting from the base population derived from the decennial Census. Thus, the greater the length of time between the mid-year estimate and the Census, the larger the number of components contributing to the uncertainty in the estimate.

Rationale for work

The UK Code of Practice for Official Statistics (UK Statistics Authority, 2008) requires that users are informed about the quality of statistical outputs. The Office for National Statistics (ONS) initiated research in 2007 to improve the understanding, measurement and reporting of the quality of the mid-year population estimates in order to address the code requirement and highlight areas where further methodological changes have the potential to achieve the greatest quality improvements.

The complexity of the problem

Measuring uncertainty in the mid-year population estimates for local authorities is complex not only due to the number of components included in each estimate, but also because the methods used to estimate these components incorporate a number of different data sources and estimation procedures. Whilst the theory of propagating variance from survey sources is well established, the components of change (births, deaths, internal migration etc.) used to compile population estimates are derived from different types of data including census, administrative and survey sources. These data sources are subject to sampling and non-sampling errors, including capture and recording errors, coverage, timeliness and inter-censual drift. For many sources, comprehensive and quantifiable information on non-sampling errors are not available. In addition data sources that provide good comparisons for the population estimates or related components are used in the estimation process itself.

Overall methodology

An evidence-based approach is used to determine error distributions for key components of error in the mid-year population estimates. A simulation based methodology is then used to produce the quality indicators. This overall approach can be broken down into several stages:

1) Initial assessment of quality issues

- Map out the procedures and data sources used to derive the local authority population estimates

[1] There are 348 local authorities (in England and Wales) with an average population of 157 thousand (mid-2009). The smallest local authority has a population of about 2,000 while the largest has a population in excess of 1 million.
[2] An internal migrant is someone who moves within the UK (England, Scotland, Wales and Northern Ireland) while an international migrant is someone who moves into or out of the UK for 12 months of more.
[3] Switchers are those who change their intentions and switch from being a migrant to being a visitor (or vice versa).

- Identify the associated quality issues
- Identify the importance of these issues by considering the likelihood of them occurring and the magnitude of the impact on the final estimate

2) Detailed investigation of quality issues

- Describe the quality issue and quantify where possible using statistical theory, empirical evidence and/or expert opinion
- Consider both sampling and non-sampling errors

3) Combine these into an overall measure of quality

It is complicated to mathematically measure the impact of a combination of possible errors to produce a final confidence interval for the population of a local authority. The key problem is that potential error distributions may be correlated and will not all be normally distributed. An alternative approach is to use a simulation based method. This would replicate the population estimates process, allowing potential errors to be calculated at each step in the process from any statistical distribution and accounting for any relationship between errors at one step and errors estimated at earlier steps.

Having derived the error distributions by local authority for each component of change for each year, values associated with these can be randomly generated using the error distributions. Conditions can be imposed on the values generated from the error distributions to allow for correlations, either across time or between components. To combine the errors for each local authority, the process of estimating the population estimates for the current year, starting with the mid-2001 base, is then simulated replacing the value of each component at each stage with the randomly generated error. For example,

Potential Error for Mid-2010 local authority estimate
= Mid-2001 error value
+ Births 01/02 error value - Deaths 01/02 error value
+ Internal In-Migrants 01/02 error value - Internal Out-Migrants 01/02 error value + ...
+ Births 02/03 error value ….

Note that the estimated error values for each component can be positive or negative. A positive error value for the deaths estimates would decrease the local authority population estimate. Hence it is subtracted in the above equation. Alternatively, a negative error for the number of deaths would increase the population size. This would create a minus minus (plus) value in the above equation.

This simulation is then repeated a large number of times (say 5000) to produce a plausible range of composite error estimates for each local authority. A composite quality measure can then be derived from the resulting distribution e.g. a 95 per cent confidence interval. In practice, it would be misleading to quote exact error values from this methodology because with non-sampling error there will always be some degree of uncertainty in these measures of uncertainty. Instead, the aim is to produce an approximate indicator of quality for each local authority estimate.

Initial work to demonstrate the feasibility of a simulation approach was carried out. For this work, fairly simplistic potential uncertainty ranges and distributions were attributed to each component. This preliminary work highlighted the existence of conflicting opinion between experts on both the distribution and magnitude of sources of error, so alternative sets of assumptions were considered.

The results showed that the components with the largest overall impact were the mid-2001 base, the internal migration estimates, and the international migration and visitor switcher components derived from the International Passenger Survey (IPS). Reflecting local population characteristics, other components such as armed forces were important in specific local authorities. The contribution of individual components was related to both the subjective error distribution assumed and the size of the component in the local authority. Internal and international migration were prioritised for further investigation.

Internal migration

Internal migration estimates are based on change of address as recorded on doctors' records (GP re-registrations). A time lag of one month between moving and re-registering is assumed. The National Health Service Central Register (NHSCR) captures all moves down to former health authority level. Below this level, moves to and from each local authority are estimated by comparing address data from annual downloads from the Patient Registration Data Service (PRDS). As this comparison will not capture multiple moves during the year or moves by people who are not included in one or other of the downloads (e.g. 0 year olds, recent immigrants and emigrants), the estimates are constrained upwards to the more complete data from the NHSCR.

Initial work focused on producing error distributions for individual quality issues and combining these to produce an overall error distribution for internal migration[4]. However this work was very time consuming, so a review of the methodology was undertaken (Smith P W F et al. 2010). An alternative approach was proposed which compares the internal migration estimates (for a given year, say 2001) to some external benchmark (such as the 2001 Census) and then fits a regression model to the log of the scaling factors for local authorities. This regression model can then be applied to subsequent years. This approach should be more efficient as it estimates the uncertainty due to several quality issues at the same time[5]. However, it is very dependent on the accuracy of the benchmark and has the weakness that the underlying model is based on a single point in time, making it difficult to capture accurately any change in the quality of the estimates over time. As well as considering the accuracy of the benchmark, it is important to assess its comparability with the internal migration estimates in terms of coverage and definitions. The following issues were identified:

- Migration flows refer to slightly different time periods (year to 29 April in the 2001 Census)
- Different population bases for migration
  e.g. Census migration estimates include moves of armed forces
- Different coverage/definition of migration moves
  e.g. a migrant in the Census is a person with a different usual address one year ago, so a person with the same address 12 months ago but a different address 6 months ago will not be included as a migrant

To improve comparability, Census data were obtained which matched as closely as possible the population covered by the internal migration estimates. The definitional issues were addressed by using the unconstrained estimates from the PRDS. These unconstrained estimates, like the Census estimates, identify migrants as people who had a different address one year ago.

The calculated scaling factors (Census/PRDS) for 2001 vary by age, sex and local authority, Figure 1 illustrates this for age and sex, while Figure 2 illustrates this for local authorities. A value greater than zero indicates that fewer migrants have been captured in the PRDS data than in the Census data. In Figure 1, larger scaling factors are seen for males than for females, and also for migrants in their twenties and thirties suggesting that migration moves are being under-estimated for these groups, which is consistent with previous research. For younger ages, the negative scaling factors indicate an overcount of migrants on the PRDS. This is due to some moves of school boarders being captured by the PRDS and these have been removed from the Census data. As the population estimates process adjusts for school boarders separately these moves will be double-counted.

In Figure 2 the local authorities have been clustered into four groups with similar age-sex scaling factor profiles. This is an initial clustering which has since been refined. Cluster 1 have been identified as predominately rural local authorities, with cluster 2 identified as predominately local authorities with boarding schools (illustrating the double counting of school boarders), cluster 3 is predominately urban local authorities and cluster 4 is predominately London local authorities and those with armed forces populations.

---

[4] Six key quality issues were identified and error distributions were produced for two issues:
- Evidence of longer or differential time lags (by age and sex) between moving and reregistering
- Double counting of school boarder moves (in both the internal migration estimates and the school boarder adjustment).

Full details of the research can be found in '*Measuring Uncertainty in the Local Authority Population Estimates: Interim Report focusing on Internal Migration*' (ONS 2009c).

[5] This approach accounts for three of the six quality issues identified for internal migration including uncertainty due to time lags, school boarders and any differences between the a person's location (or inclusion) in the GP data and population estimates in 2001 (using 2000 as a proxy). It will not cover, for example, errors associated with constraining the PRDS estimates to NHSCR data.

Figure 1: Mean log scaling factors (across all local authorities) for inflows by age and sex

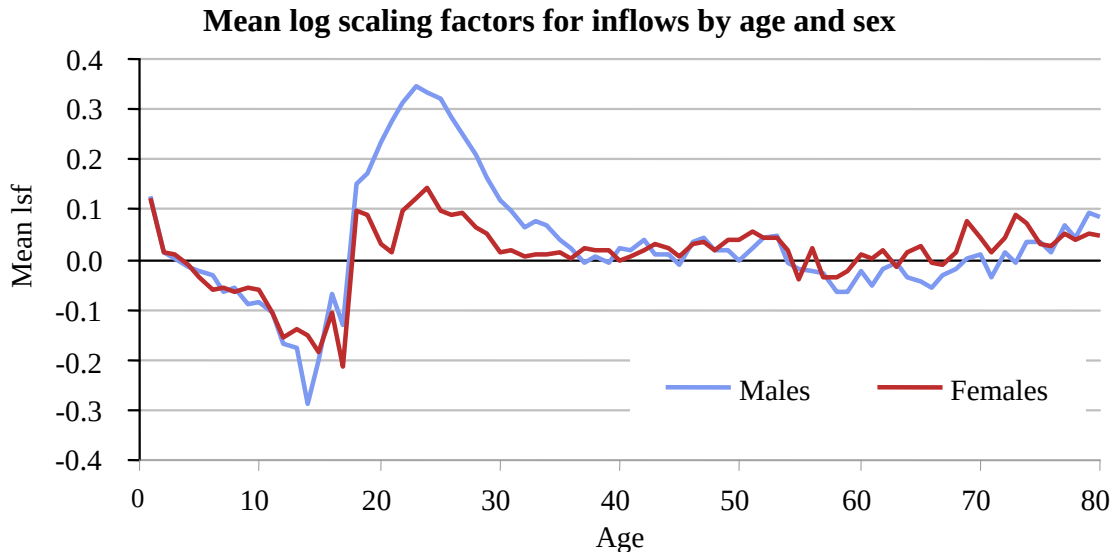
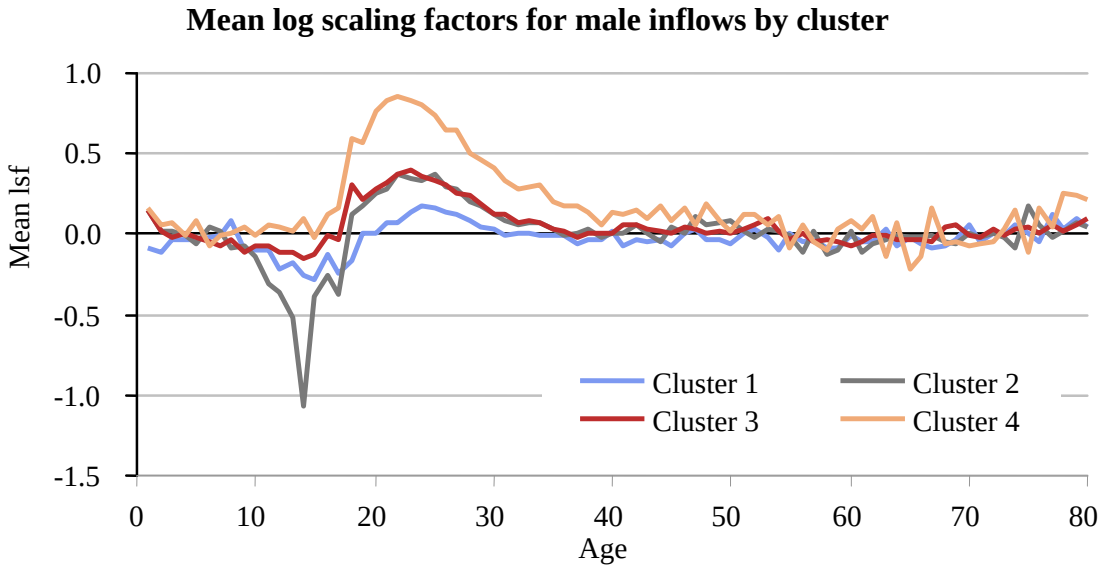


Figure 2: Mean log scaling factors for inflows for each cluster by age (males)



To estimate the error distributions, a non-linear regression model has been fitted to the log of the scaling ctors, where the variation by age is captured using Rogers-Castro curves. Separate curves are fitted for inflows and outflows, for males and females and for different clusters of local authorities. A step change in the 'overall level' parameter has been added, so that the level of the curve can be lower for younger ages (<18) than for older ages. Additional covariates have been added for school boarder (12-17) and student ages (18-19) so that some of the parameters become functions of covariates. This allows variation by local authorities within each cluster. The form of the model for *inflows* for a given sex *y* and cluster *k* is:

$$\log\left(\tilde{C}_{ix} / \hat{C}_{ix}\right) = a_6 - a_7 \sum_{j=0}^{17} \mathrm{I}_j(x) + \sum_{j=12}^{17} b_j u_{ij} \mathrm{I}_j(x) + \sum_{j=18}^{19} c_j v_{ij} \mathrm{I}_j(x) + a_0 \, \mathrm{e}^{-a_1 x} + a_2 \, \mathrm{e}^{-a_3(x-a_4)-\mathrm{e}^{-a_5(x-a_4)}} + \varepsilon_{ix}$$

where

$\tilde{C}_{ix}$ = the (adjusted) 2001 Census internal migration inflow for local authority *i* (in cluster *k*), sex *y* and age *x*

$\hat{C}_{ix}$ = the PRDS internal migration estimate for local authority *i* (in cluster *k*), sex *y* and age *x*

$\mathrm{I}_j(x) = 1$ if $x = j$, 0 otherwise

u_{ij} = the number of school boarders aged *j* in local authority *i* (j = 12 to 17)

v_{ij} = the number of students aged *j* in local authority *i* (j = 18 to 19)

$a_0, a_1, ..., a_7 ; b_{12}, b_{13}, ..., b_{17} ; c_{18}, c_{19}$ are the estimated parameters for sex *y* and cluster *k*

ε_{ix} are independent, normally distributed error terms with mean 0 and variance σ^2

Additional covariates may be added for other ages and/or other parameters depending on the fit of the model. A distribution of scaling factors for each age-sex group within each local authority can then be simulated from the models by re-sampling from the distribution[6] of ε_{ix} and adding the re-sampled value to the predicted value from the model. To convert these scaling factor distributions into error distributions, the scaling factors are applied to the internal migration estimates for 2001.

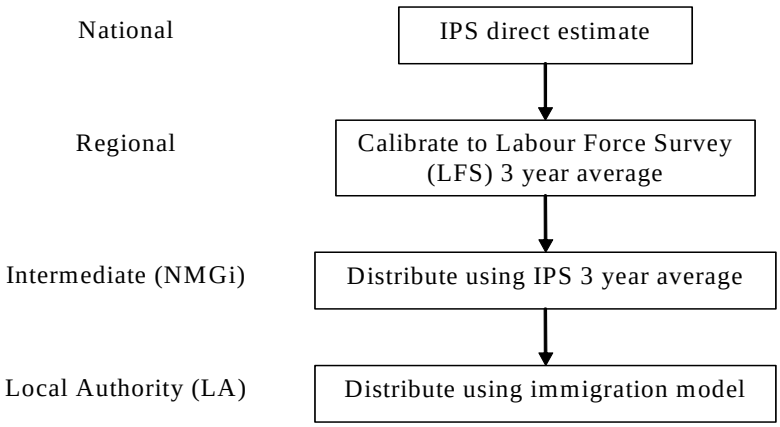
To obtain error distributions for other years, a way of measuring change over time has to be developed. The simplest option is to assume that the scaling factors or relative error (by age, sex and local authority) remains constant over time or retains the same relationship with the covariates (if covariates are included in the model). Alternatively, it can be assumed that the relationship for each cluster remains constant but that local authorities may change cluster. This latter approach depends on being able to predict cluster membership based on external factors which can be captured through annually updated covariates.

International migration

For international migration, a methodology is being developed to measure IPS statistical error for local authorities. The term 'statistical error' relates to sampling error introduced through the use of survey estimates and also error in predicted values derived from the statistical models used at local authority level.

The 'IPS flow' is the largest component of the international migration estimates. It estimates the number of people going to or coming from an area who intend to be long-term migrants[7]. A multi-stage approach is used to produce the IPS estimates at local authority level. This is illustrated (in outline) for immigration for local authorities outside the London region[8] in Figure 3. Further details of the methodology and the equivalent methodology for emigration can be found in '*Estimating international long-term immigration by local authority'* and '*Estimating international long-term emigration by local authority'* (ONS 2009a, b).

Figure 3: Method for distributing immigration to local authority level outside the London region[8,9]



To estimate IPS statistical error three different methods were proposed by Smith P W F, Bijak J, Raymer J (2010): benchmarking, simulation and an analytical approach. Both the simulation and analytical approaches have the advantage that they can use up-to-date information on the sampling error in both the Labour Force Survey (LFS) and IPS, and capture any change in the precision of the models over time. While the 2001 Census could potentially be used as a benchmark for immigration, no benchmark exists for emigration. A population accounting approach (which is sometimes used to produce a benchmark for emigration) has the major flaw that the resulting benchmark will be affected by any errors in the other components of the equation (births, deaths, internal migration and changes in special populations) which may be greater than the uncertainty being

[6] The error terms ε_{ix} (for a given sex *y* and cluster *k*) could either be simulated from an assumed normal distribution with mean 0 and variance $\hat{\sigma}^2$ (estimated from the data) or sampled with replacement from the empirical distribution of the residuals.

[7] Respondents in the IPS are interviewed at their point of arrival (or departure) and so are asked what they plan (or intend) to do. Some migrants (and non-migrants) will change their intentions and hence their status. Adjustments for these 'switchers' are included as separate components in the international migration estimates.

[8] For London, the methodology is different: after calibration to the LFS, the regional total is split into students and non-students. Students are distributed to NMGi's using Census data and non-students using the LFS, before being re-combined.

[9] Calibration to the LFS data (which records actual destination) at regional level removes geographical bias through the use of 'intentions based' IPS data. Modelling is used at LA level as the IPS sample size is insufficient.

estimated. Some other reasons why a benchmark approach was not used for immigration are:

- The immigration estimates derived from the Census are on a different definition to the IPS component and may include some short-term migrants.
- The benchmark approach assumes that the relative error in the estimates does not change over time which is unlikely to be true due to recent improvements in the IPS and changes in the pattern of international migration due to events such as EU Accession.
- Immigration estimates derived from the Census include some migrant subgroups[10] which are not in the IPS component, and it would not be possible to remove all of them from the data.

The main advantage of the benchmark approach is that it captures both sampling and non-sampling error.

The analytical approach was also not taken forward. While in theory it should be possible to derive an equation for statistical error in the local authority estimates, previous research has shown that this is computationally complex, in particular due to the use of multiple constraining. Another disadvantage with the analytical approach is that it is very dependent on the underlying assumptions which might not hold.

Consequently, we decided to use a simulation approach. This requires creating a number of simulated data sets for the LFS and IPS which reflect the sampling error in the original samples. Both parametric and non-parametric bootstrapping are used in the simulation methodology. Parametric bootstrapping (used to simulate the non-zero local authority IPS estimates) assumes an underlying distribution, and simulates from this distribution, while non-parametric bootstrapping (used to simulate the LFS data) makes no distributional assumption and re-samples (with replacement) directly from the observed data. When re-sampling the LFS, we have chosen to replicate each observation in proportion to its weight, and then draw a simple random sample (with replacement) from this pseudo population.

One of the problems encountered in implementing the methodology was how to replicate the large number of zero observations in the IPS at local authority level. These observations, unlike the rest of the survey data have no associated standard errors and vary in location from year to year (and from sample to sample). We plan to simulate the IPS data in two-stages. First we will use a logistic model (based on historical data) to decide if the estimate is zero or non-zero. A second model will then determine the estimate for the local authority.

Each of these simulated data sets will then be used to produce a set of local authority IPS estimates using the immigration methodology. For each local authority, the difference between the simulated estimates and the original estimates will be calculated and these differences will be used to produce an error distribution for each local authority. The simulation approach has the advantage that it is general and flexible, does not require a benchmark, and unlike the analytical method is less reliant on model assumptions.

Next steps

Using the estimated error distributions (by local authority) for internal migration and the IPS statistical error, errors will be randomly generated for each year and combined with simulated errors for the 2001 Census base (based on the One Number Census (ONC) standard errors) to produce an overall error distribution for each local authority for each year. Local authorities will then be grouped using these error distributions or a summary quality measure derived from these distributions, such as a 95 per cent confidence interval. This will provide a measure of uncertainty for each local authority encompassing the quality issues considered.

REFERENCES

Office for National Statistics (2009a): *Estimating international long-term immigration by local authority (LA).*

Office for National Statistics (2009b): *Estimating international long-term emigration by local authority (LA).*

Office for National Statistics (2009c): *Measuring Uncertainty in the Local Authority Population Estimates: Interim Report focusing on Internal Migration.*

Smith P W F, Bijak J, Raymer J, Forster J (2010): Review of ONS Proposed Methodology to Measure Uncertainty in Local Authority Populations Based on the Demographic Components of Change by Age and Sex, University of Southampton.

Smith P W F, Bijak J, Raymer J (2010): Report on Developing a Methodology to Estimate Statistical Error in Local Authority International Passenger Survey Estimates of Immigration and Emigration.

UK Statistics Authority (2008): *Code of Practice for Official Statistics.*

[10] Examples include asylum seekers and their dependents, visitor switchers and armed forces.