2011 UK Census Coverage Assessment and Adjustment Methodology

Owen Abbott, Office for National Statistics, UK[1]

1. Introduction

The central objective of the 2011 UK Census is to provide high quality population statistics as required by key users such as policy makers and service providers. It is being designed to count everyone and every effort will be made to ensure that everyone is counted, and so there are a number of new innovations in the design of the 2011 Census. For the first time (in England and Wales), questionnaires will be sent by post, which will rely on the construction of a reliable household frame and a robust publicity campaign. This allows resources to be redirected into the follow-up operation, where the field force will be flexible in order to be able to react quickly to areas of poor response. Finally, respondents will be able to respond over the internet.

Whilst these measures are designed to ensure a high response, not everyone will be counted in the 2011 Census. This undercount does not occur uniformly across all geographical areas or across other sub-groups of the population such as age and sex groups. In terms of resource allocation, this is a big issue since the population that is missed can be the part which attracts higher levels of funding. ONS outlined its coverage assessment and adjustment strategy in Abbott (2007). This paper outlines the proposed methods for producing the 2011 Census estimates.

The methodology builds on the 2001 One Number Census (ONC) – this used a large survey in combination with the 2001 Census to estimate coverage. This is a standard technique for measuring the quality of a census, used by many National Statistics Institutes. The Census outputs are adjusted prior to their release to include those individuals and households estimated to have been missed. The methods used in adjustment are established statistical techniques based on imputation.

In 2001, for some Local Authorities (LAs - the key local government unit to which central funds are distributed) this meant that their Census results included more than 20,000 individuals who had not been counted, but had been estimated. It is important that users of census data understand that the Census population numbers are estimates (not counts), based on a combination of the Census and the Census Coverage Survey (CCS).

The methodology is described in greater detail in Census Advisory Group paper (08)05 available at http://www.statistics.gov.uk/census/pdfs/ag0805.pdf

2. Background - The 2001 One Number Census

Measured census undercount levels have on the whole been increasing over the past few decades. The differential nature of the undercount is important since, for

[1] Address for correspondence: Owen Abbott, Office for National Statistics, Room 4200N, ONS, Segensworth Road, Titchfield, Fareham, PO15 5RR, UK

example, young males in inner city areas are difficult to enumerate. This has led to increasing focus on the methods for measuring this differential undercount.

In the 2001 UK Census, the One Number Census (ONC) project estimated and adjusted for the number of people and households not counted in the 2001 Census. The methodology is described by Brown *et al* (1999). The aim was to provide one number – the national population estimate – to which all census tabulations would add up. The ONC measured the undercount in the 2001 UK Census to be 6 per cent of the total population (approximately 3.1 million individuals). This means that the Census achieved an estimated 94% response (94% in England and Wales, 96% in Scotland and 95% in Northern Ireland). This compares to response rates of 98% in Australias 2006 Census and 97% in the 2001 Canadian Census.

In the 2001 Census, the response varied widely across Local Authorities (LAs) – the lowest was 64% and there were around 30 LAs (out of 376) with a response rate lower than 90%. There were some issues with the results which led to further studies and adjustments – there were some areas where local Census failures resulted in estimates that were too low. The lessons are summarised by ONS (2005). However, in the majority of Local Authorities the results were of high quality.

3. 2011 UK Coverage assessment and adjustment strategy

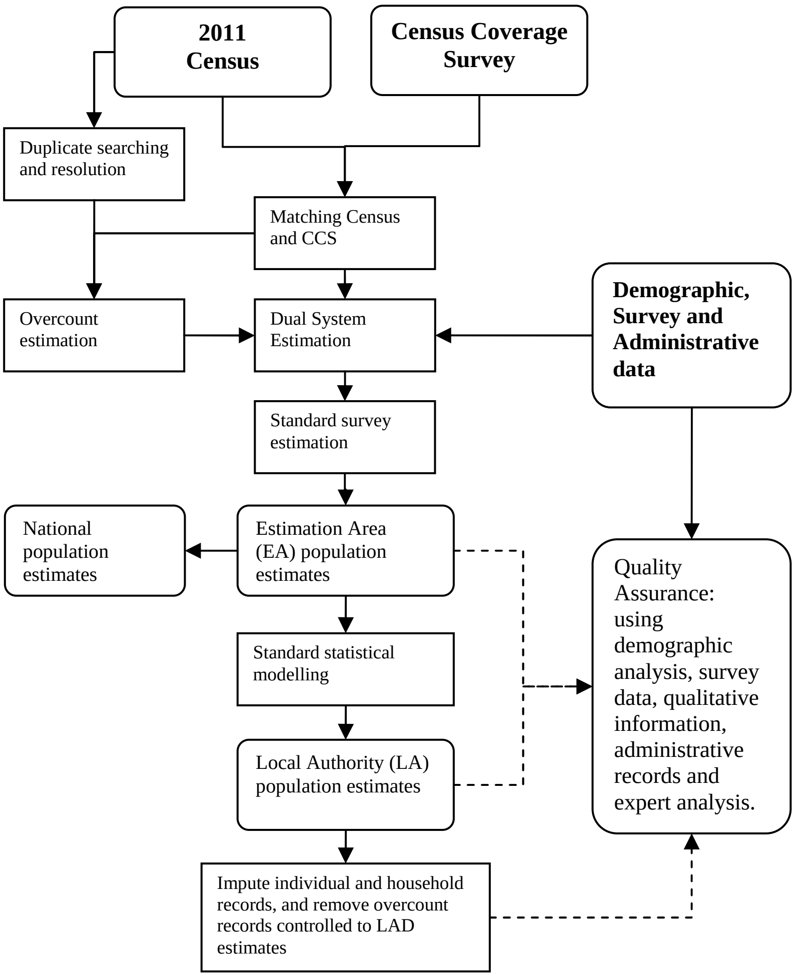
The primary objective of the coverage assessment and adjustment strategy in 2011 is to identify and adjust for the number of people and households not counted in the 2011 Census. A secondary objective is to identify and adjust for the number of people and households counted more than once, or counted in the wrong place, in the 2011 Census. The ONC framework will be built upon, using it as a platform to develop an improved methodology.

Given that the estimates will be based on a survey, it is important to be able to quantify the accuracy component of the quality of the estimates by providing confidence intervals. The target levels for these are relative confidence intervals of 0.2 per cent around the national population estimate (i.e. plus or minus 120,000 persons) and 2 per cent for a population of half a million. This means that the national population estimate will have greater relative precision than regional or local estimates. The achieved levels will depend on a number of factors including CCS sample size, Census response and CCS response.

4. Summary of Methodology

The methodology being proposed to achieve the above strategic aims and objectives is described in the following sections. The key stages are shown in Figure 1.

Figure 1 – The 2011 Coverage Assessment and Adjustment process overview


2011 Census
Census Coverage Survey
Duplicate searching and resolution
Matching Census and CCS
Overcount estimation
Dual System Estimation
Demographic, Survey and Administrative data
Standard survey estimation
National population estimates
Estimation Area (EA) population estimates
Quality Assurance: using demographic analysis, survey data, qualitative information, administrative records and expert analysis.
Standard statistical modelling
Local Authority (LA) population estimates
Impute individual and household records, and remove overcount records controlled to LAD estimates

The methodology can be summarised as follows:

a) A Census Coverage Survey (CCS) will be undertaken, independently of the Census. The survey will be designed to establish the coverage of the Census. A sample will be drawn from each Local Authority.

b) The CCS records are matched with those from the Census using a combination of automated and clerical matching.

c) The census database is searched for duplicates, and, together with the CCS, the levels of overcount in the census are estimated.

d) The Census estimates are produced within groups of similar Local Authorities (called Estimation Areas (EAs)) using standard statistical techniques.

e) Standard statistical modelling methods will then be used to estimate the Local Authority (and possibly lower level) Census estimates.

f) Households and individuals estimated to have been missed from the Census will be imputed onto the Census database, with allowance for overcount. These adjustments will be constrained to the LA level estimates.

g) All the Census estimates are quality assured using demographic analysis, survey data, qualitative information and administrative data to ensure the estimates are plausible.

5. The Census Coverage Survey

The key element in the coverage assessment and adjustment methodology is the CCS. A representative sample of around 16,500 postcodes (units used by the mail system), which is about 320,000 households, will be surveyed in England and Wales. It will be undertaken during a four week period starting 6 weeks after Census Day and will be operationally independent of the Census enumeration. A short, paper-based interviewer-completed questionnaire will be used (as opposed to the Census self-completion questionnaire) designed to minimise the burden on the public. The CCS fieldwork will be very similar to that employed for the 2001 CCS, as the survey was broadly a success (see Abbott *et al*, 2005).

The survey will be designed to enable the Census to be adjusted for undercount at the national, local authority and smaller area level. The sample will be area based to enable both coverage of households and individuals within households to be measured. A sample of postcodes will be drawn from each Local Authority and all households within the sampled postcodes will be interviewed.

The sample design strategy will spread the sample across different area types based on a national Hard to Count (HtC) index created from variables most associated with undercount in the 2001 Census and those believed to be relevant in 2011. Listed in order of importance these are households:

- renting privately;
- where the occupants are of Black, Asian, Chinese or Mixed ethnic group;
- paying part rent/part mortgage;
- containing a single person; and;
- where the average age of the people within the household is between 23 and 34.

The most up-to-date data sources will be used where possible to ensure that the sample includes areas that are both easy and difficult, particularly in local areas that are changing over time.

6. Matching and Estimation

The Census estimates will be based on a methodology known as Dual System Estimation (DSE). It is inevitable that some households and people will be missed by both the Census and CCS but DSE can be used to make an adjustment for this by considering the relative numbers of the people observed by:

- both the Census and CCS;
- the Census but not the CCS; and
- the CCS but not the Census.

In order to identify the numbers in each of these groups, the records from the CCS must be matched against those from the Census. It is essential that this matching process is accurate as missed matches will create an overestimate of the population.

The 2011 matching strategy will be similar to that developed in 2001 by Baxter (1998), involving a combination of automated and clerical matching. Both exact and probability matching techniques are used, which are standard ways of maximising the automatic match rate. The matching strategy will be designed to minimise the missed match rate and so significant clerical resource will be required.

The next stage is to estimate the undercount for all LAs using the combined Census and CCS data generated by the matching. There are three stages in the process.

6.1 Stage 1 – Dual System Estimation (DSE) within sampled areas

This output from the matching process will be used to estimate the undercount for each CCS postcode, stratified by age and sex using DSE, a technique applied widely and also known as capture-recapture. After matching between the Census and the CCS, we observe the proportion of individuals captured by the CCS that were also captured by the Census. We then assume that this same proportion holds for the total people counted by the Census – so the total population is then the Census count multiplied by the inverse of the probability of being able to match an individual in the CCS to the Census (or the ratio of the CCS count to the matched count).

This technique effectively estimates the number of persons missed by both the Census and CCS. However, DSE does require some conditions to be met and violation of these will result in Census estimates that are too low or too high. In the 2001 ONC process, quality assurance showed that there was some under-estimation and as a result, Brown *et al* (2006) developed a method to make adjustments to the DSEs by incorporating additional data (which added 230,000 to the national Census estimate). For 2011, correcting for such problems in the DSE will be a part of the methodology. This will use a similar approach to that used in 2001, which used a statistical model and a household count to make an adjustment. However, in 2011 it will be designed to include additional reliable sources of data, such as demographic sex ratios or administrative sources.

6.2 Stage 2 – Estimation Area estimation

The second stage in the estimation process is to generalise the DSEs to the non-sampled areas. The sample is pooled into Estimation Areas (EAs) to ensure that sample sizes are adequate for robust estimation. Estimation Areas are groups of LAs with similar coverage patterns. Similar LAs will be grouped together using a national classification to form the EAs, which will contain half a million individuals on average. This means that estimates are produced for some large LAs individually.

Standard survey estimation methods will be used for the generalisation – in its simplest form this will estimate the average undercount seen in the DSE across the sampled areas (which is about 1% of postcodes) and then apply that to the non-sampled areas (the other 99% of postcodes).

The output from this will be the Census estimates for each EA by age and sex, together with an indication of their accuracy. All of the subsequent stages will be consistent with these Census estimates – they are the ‘best’ adjusted Census estimates of population.

6.3 Stage 3 – Local Authority District Estimation

Since many EAs will consist of more than one LA, estimates of the age-sex population for each LA will need to be made. Many of the LAs are unlikely to contain sufficient CCS postcodes to enable accurate estimates of population to be made with the sample from the LA alone even though the sample will cover all LAs. Statistical modelling techniques can be applied to produce more precise LA level population estimates using the sample information from other LAs in the same Estimation Area. This will mean better accuracy than if the sample from the LA was used in isolation.

The modelled LA level estimates will be adjusted so that they sum to the EA level estimates, and their accuracy can also be calculated to provide 95% confidence intervals around the LA Census population estimates.

7. Measuring overcount

The 2001 One Number Census focused on measuring the population by adjusting for undercount. Overcount has not historically been a problem within the UK Censuses, and therefore measurement of it was given a low priority. Studies of duplicates within the 2001 Census estimated that there were potentially around 0.4 per cent (around 200,000) duplicate persons. However, no adjustments were made.

One of the improvements for 2011 is a more rigorous measurement of overcount. Abbott and Brown (2007) present a full discussion of the options for measuring overcount within the existing framework. A number of sources of information could be used to estimate the level of overcount, including searching the database for duplicates and using the CCS to detect individuals who are counted in the wrong location. The Census estimates will be adjusted to take account of the estimated level of overcount. This means that in some LAs where the estimated overcount is high their Census estimates will be reduced to take this into account. However, in general it is anticipated that the levels of overcount will be much lower than the levels of undercount.

8. Adjustment

The final stage prior to Quality Assurance is the creation of an adjusted census database that is consistent with the Census estimates. This will use a similar methodology to that used in 2001, described by Steele *et al* (2002), albeit with improvements designed to provide more robust results. The information on the characteristics of missed persons obtained in the CCS will allow the creation of this database. Wholly missed households will be imputed, located using the Census household frame, and persons within counted households will also be imputed to account for those missed by the Census.

The result is an individual level database that represents the best estimate of what would have been collected had the 2011 Census not been subject to undercount or overcount. This database will be used to generate all statistical output from the Census, and so all tabulations will automatically include compensation for coverage errors for all variables and all levels of geography, and will be consistent with the Census estimates.

9. Quality Assurance

A quality assurance process will be undertaken to ensure that the Census estimates are plausible and of the right overall magnitude. This will involve a series of aggregate level quality checks, aided by data, grouped by age, sex, other important variables and geography. The strategy is likely to be similar to the model used in 2001 (described in ONS, 2005), expanded to include more data sources and more comparisons. The types of sources that could be used in the Quality Assurance process are:

- Demographic mid-year population estimates;
- Numbers of people listed on health registers;
- Social security information;
- Education information;
- Estimates of population characteristics from large surveys;
- Information from Longitudinal Studies
- Visitor data collected in the Census; and;
- Demographic analyses (such as Sex or mortality ratios);

In addition, a range of descriptive information will be gathered to give a fuller picture of the area under consideration, such as management information from the census processing operation or intelligence gathered on the data sources.

A panel consisting of specialists will consider the evidence for each LA before either accepting or rejecting the estimates. Contingency strategies will be developed and used if initial estimates are rejected. This might include a strategy that uses a plausible target sex ratio to estimate the young male population, assuming the estimates of young females are correct. The QA process will also include consideration of regional, national and special population estimates.

10. Summary

The 2011 Census project has a number of initiatives to improve the enumeration process and deliver a high quality census. Despite these efforts, the 2011 Census will both miss people and also count them more than once. Evaluation of such coverage errors is critical. The framework for measuring coverage of the Census is based on that developed for the 2001 One Number Census and its Census Coverage Survey. For the 2011 Census, there are a number of improvements being explored within this framework in order to deliver high quality Census estimates. These improvements will help to ensure that users of the 2011 Census data are confident in the results.

References

Abbott, O., Jones, J. and Pereira, R. (2005) “2001 Census Coverage Survey: Review and Evaluation”, *Survey Methodology Bulletin,* Newport, UK: Office for National Statistics*.* 55, pp. 37-47.

Abbott, O. and Brown, J. (2007) “Overcoverage in the 2011 UK Census”, 2007 Proceedings of the American Statistical Association, Survey Research Section [CD-ROM], American Statistical Association, Alexandria, VA.

Abbott, O. (2007) “2011 UK Census Coverage assessment and adjustment strategy”. *Population Trends*, 127, pp. 7-14. Available at www.statistics.gov.uk/downloads/theme_population/PopulationTrends127.pdf

Baxter, J. (1998) “One Number Census matching”. One Number Census Steering Committee paper 98/14. Available at www.statistics.gov.uk/census2001/pdfs/sc9814.pdf

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999) “A methodological strategy for a one-number census in the UK”. *J. R. Statist. Soc. A*, 162, pp. 247-267.

Brown, J., Abbott, O., and Diamond I. (2006) “Dependence in the one-number census project”. *J. R. Statist. Soc. A*, 169, pp883-902.

ONS (2005) “One Number Census Evaluation Report”. Available at www.statistics.gov.uk/census2001/pdfs/onc_evr_rep.pdf

Steele, F., Brown, J. and Chambers, R. (2002) “A controlled donor imputation system for a one-number census”. J. R. Statist. Soc. A, 165, pp. 495-522.