2011 UK Census Coverage Assessment and Adjustment Methodology

Owen Abbott
Office for National Statistics

Every effort is made to ensure everyone is counted in a census. However, no census is perfect and some people are missed. This undercount does not usually occur uniformly across all geographical areas or across sub-groups of the population such as age and gender. Further, the measurement of small populations, one of the key reasons for carrying out a census, is becoming increasingly difficult. In terms of resource allocation, this is a big issue since the people that are missed can be those who attract higher levels of funding. Therefore money may be wrongly allocated if the Census is unadjusted. ONS outlined its coverage assessment and adjustment strategy in Population Trends 127 (see Abbott, 2007), noting where improvements over the methodology used in 2001 would be sought. This article outlines the proposed methodology for the 2011 Census arising from that strategy, and focuses on the research that has been conducted to date to develop those improvements and innovations.

Introduction

The census provides a once-in-a decade opportunity to get an accurate, comprehensive and consistent picture of the most valuable resource of the UK – its population – and a rich array of facts about it (Cabinet Office, 2008). The key strategic aims include:

- giving the highest priority to getting the national and local population counts right
- maximising overall response and minimising differences in response rates in specific areas and among particular population subgroups
- provision of high quality, value-for-money, fit-for purpose statistics that meet user needs and which are as consistent, comparable and accessible across the UK as is possible

It is widely accepted practice that when conducting a traditional style census, an assessment of coverage should be part of the statistical operation. The UK is no exception, and the 2001 Census represented the first real attempt to fully integrate the census and coverage measurement processes, resulting in the development of the One Number Census (ONC) methodology (see Holt *et al*, 2001). The aim was to provide a population estimate that would be the basis for the 2001 mid-year estimate, and to which all census tabulations would add. The ONC estimated the undercount in the 2001 Census to be 6.1 per cent of the total population in England and Wales, 3.9 per cent in Scotland and 4.7 per cent in Northern Ireland.

The 2001 methodology was a big step forward. Both the Statistics Commission (2003) and the Local Government Association (2003) published reviews that concluded that the methodology used in 2001 was the best available and no alternative approach would have produced more reliable results overall. However, there were some issues with the results which led to further studies and adjustments, summarised by ONS (2005). These adjustments added another 0.5 per cent to the estimated population of England and Wales. As a result, there were a number of key lessons from the ONC project which were fully evaluated by ONS (2005). In summary, these lessons were:

- The methodology was not able to make adjustments in all situations, particularly when there were pockets of poor census response
- Engagement with stakeholders is critical
- That the methodology needs to be robust to failures in underlying assumptions and in particular have inbuilt adjustments for such failures – e.g. lack of independence between the census and the Census Coverage Survey (CCS)
- Two of the weaknesses of the methodology were not having additional sources of data to complement the CCS, and the perception that it would solve all 'missing data' problems
- The measurement of overcount requires greater attention
- The balance of 'measurement' resource between easier and harder areas needs careful consideration, as more sample in harder areas may even out the quality of the estimates

This article provides a summary of the high level strategy described by Abbott (2007) and then outlines the methodological framework. The detailed methodology for each of the components is summarised, including the design of the coverage survey, the estimation process and the improvements that have been introduced.

This article is in the main about the methodology as it applies to England and Wales. However, although the methodology is applicable to the UK, it is expected that there will be slight differences between countries to reflect local circumstances. The differences have not been highlighted in this article.

2011 Coverage assessment and adjustment strategy

As outlined in Abbott (2007), the coverage assessment and adjustment strategy in 2011 is to develop an improved methodology built on the 2001 framework. The improvements sought are closely linked to the data and lessons learnt from the 2001 experience as well as anticipated changes to the population and census methodology over the intervening decade.

There are a number of other objectives, summarised in **Box one**.

Methodology

The methodology used to achieve the strategic aims and objectives is described in the following sections. The key stages are shown in **Figure 1**, and can be summarised as follows:

(a) A CCS will be undertaken, independently of the census. The survey will be designed to estimate the coverage of the census. A sample will be drawn from each local authority (LA).
(b) The CCS records are matched with those from the Census using a combination of automated and clerical matching.
(c) A large sample of census records are checked to see if they are duplicates. The CCS is then used to help estimate the levels of overcount in the census, by broad age-sex groups and Government Office Region.
(d) The undercount is estimated within groups of similar LAs (called Estimation Areas (EAs)) to ensure that sample sizes are adequate. The matched Census and CCS data are used within a dual system estimator (DSE), which is augmented with other reliable sources of data such as the census household frame to estimate and adjust for any residual bias. These DSEs are then used within a simple ratio estimator to derive undercount estimates for the whole of the Estimation Area.
(e) The population estimates for the Estimation Areas are then calculated using the undercount and overcount estimates.
(f) Small area estimation techniques will then be used to estimate the LA population estimates.
(g) Households and individuals estimated to have been missed from the census will be imputed onto the census database, after taking into account the estimated overcount. These adjustments will be constrained to the LA estimates.
(h) All the population estimates are quality assured using demographic analysis, survey data, census information on visitors, qualitative information and administrative data to ensure the estimates are plausible. This component is not covered in this paper, as it is a separate and significant stream of research. This will be addressed in a future *Population Trends* article.

Box one

Summary of coverage assessment and adjustment objectives

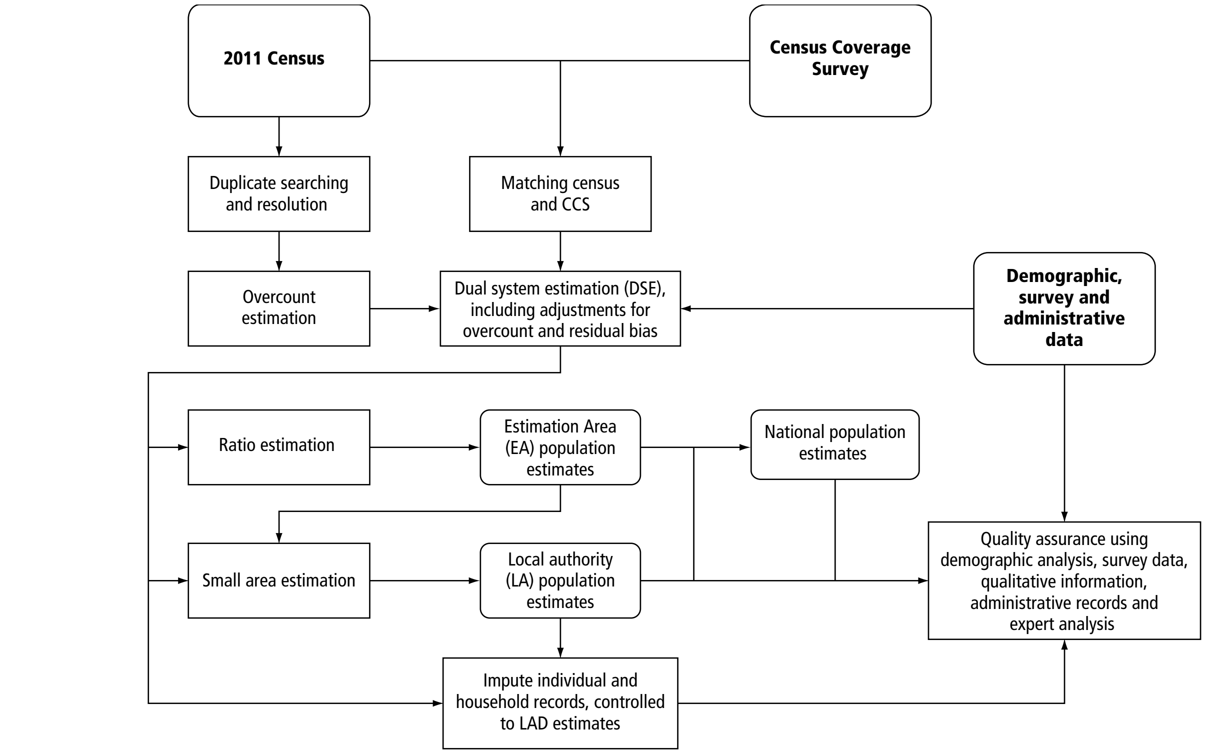
- Address the lessons from 2001, looking for improvements and taking into account the changes to the census design
- Measurement of over-coverage should be addressed
- Gaining acceptance of the methodology from users is a key objective. Users will not accept their census population estimates if they are not confident about the methodology used to derive them
- Simple methods should be developed where possible, to aid communication of the methodology
- Since all census outputs will be influenced by the methodology, we will communicate with all users through appropriate channels and with tailored materials
- There are a number of ways in which undercount can occur (such as missing a whole household or missing a person from a counted household), and an objective is to measure the extent of each of these, permitting more transparent adjustments
- Aim for the local authority and age-sex level population estimates to be the same relative precision across all LAs
- Target precision rates are 95 per cent 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 (i.e. plus or minus 10,000 persons)
- Ensure that there are no LAs with a worse precision than the worst that was achieved in 2001 and improve the worst 5 per cent of areas (i.e. there is no relative confidence interval for a LA total population that is wider than 6.1 per cent, and a 5 per cent confidence interval is the desirable upper bound).

The Census Coverage Survey

The key element in the coverage assessment and adjustment methodology is the CCS. This section details the sampling methodology used, the sample size of the survey and key aspects of the survey methodology. Important features of the CCS include:

- It will be designed to enable census population counts to be adjusted for underenumeration at the national, local and small area level
- It will comprise an intensive enumeration of a representative sample of between 15,000 and 16,000 postcode units across England and Wales. The sample of postcodes will be drawn from all local

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 (DSE), including adjustments for overcount and residual bias
Demographic, survey and administrative data
Ratio estimation
Estimation Area (EA) population estimates
National population estimates
Small area estimation
Local authority (LA) population estimates
Quality assurance using demographic analysis, survey data, qualitative information, administrative records and expert analysis
Impute individual and household records, controlled to LAD estimates

authorities. The national sample size is approximately the same as was used in 2001

- It will consist of a short, paper-based interviewer-completed questionnaire (as opposed to the census self-completion questionnaire) designed to minimise the burden on the public, and therefore maximise response rates. This will be vital since the CCS, unlike the census, will be a voluntary survey
- It will be operationally independent of the census enumeration exercise
- It will be undertaken during a four week period starting six weeks after Census Day

Design

The CCS will be a stratified two-stage sample selection of postcodes that will be independently re-enumerated. The first stage will select a sample of Output Areas (OAs), stratified by local authority and a national Hard to Count (HtC) index. The second stage will then select three postcodes from within each selected Output Area. In 2001, five postcodes were selected in each primary sampling unit. We are selecting fewer postcodes in each, allowing us to spread the sample over more OAs. This reduces the clustering in the design, making it more statistically efficient, but increasing travelling costs slightly.

In 2001, the main geographical stratification in the design came from forming Estimation Areas (EAs) by grouping contiguous local authorities to create populations of around 500,000 people, and using these for sampling and estimation. However, for 2011 the strategy will be to draw the samples from LAs directly, but then to form the EAs at the estimation stage. This provides a sample that is better for making LA level estimates – either directly for large LAs, or by using small area estimation for smaller LAs. Where there is insufficient sample within an LA to estimate the population with an acceptable level of accuracy, we will post-stratify the LAs into Estimation Areas, possibly grouping them by area type indicators rather than restricting the groups by contiguity (although it is expected that the grouping will be constrained by the Government Office Region boundaries). This is expected to increase the efficiency of the estimation process, as areas with similar undercount patterns will be grouped together.

As undercount is disproportionately distributed across areas, the OAs within each LA are stratified according to a national HtC index. This index attempts to capture the non-geographical variation in undercount in a census. Research into the household characteristics most associated with undercount in the 2001 Census has been undertaken using various modelling approaches. The model that has been developed to predict the relative difficulty of enumerating an Output Area attempts to include timely data sources, including:

- The proportion of persons claiming Income Support or Jobseeker's Allowance
- A measure of the proportion of persons who are non-'White British'
- A measure of the relative house price within an LA
- A measure of dwelling density

The use of more up-to-date information should ensure the sample design is robust in areas of high change. The national HtC index is likely to partition all OAs in England and Wales into a 40 per cent, 40 per cent, 10 per cent, 8 per cent and 2 per cent distribution, which is similar to that used in 2001, but is more refined (the 2001 index had three levels with a 40 per cent, 40 per cent, 20 per cent distribution) because we have more confidence in the information about undercount patterns. The division of the top 20 per cent of OAs into three groups will mean that in most LAs there will always be around three HtC strata – in 2001 the top 20 per cent was concentrated in London and metropolitan LAs and thus only one HtC stratum was present in some LAs. The 2011 distribution will address this problem and provide a more localised index.

Sample selection from the above stratification requires a method of sample allocation across the strata. In 2001, the strategy was to use the previous census population counts as a proxy, and allocate the sample based upon the pattern of the key-age sex groups (see Brown *et al*, 1999). For 2011, the data obtained on coverage patterns from the 2001 Census provide a better proxy and can be used to allocate the sample. However, the actual 2011 coverage patterns are not always going to follow those seen in 2001, so a conservative allocation using the 2001 data will be adopted. A minimum sample size constraint will be applied which ensures representation for each LA. There will also be a maximum sample size constraint to guard against over-allocation based upon the 2001 situation. This will mean that areas which we expect to have a high undercount will have a larger sample than in 2001, and conversely there will be smaller sample sizes in high coverage areas. This meets the census objective of consistent quality of the estimates across areas.

This sample design strategy should provide an efficient and robust design that spreads the sample across different area types, achieving consistent quality of estimates across LAs.

Sample size

The sample size of the CCS must be sufficiently large to ensure that the accuracy of the population estimates is acceptable. The larger the sample size, the more accurate the population estimates; however this must be balanced against the cost, quality and practicalities of carrying out a larger CCS. Work has been undertaken to explore the precision of the estimates for different CCS sample sizes and census coverage patterns. Based on this, a sample size similar to that employed in 2001 of around 16,000 postcodes (about 1.2 per cent) or 300,000 households for England and Wales will provide an acceptable level of accuracy (relative confidence intervals of around 2–3 per cent) for populations of 500,000 (around 0.2 per cent for the national population).

Survey practicalities

The CCS fieldwork will be very similar to that employed for the 2001 CCS as described by Pereira (2002), as the survey was broadly a success (see Abbott *et al*, 2005).

Box two

Census Coverage Survey topics for the 2009 Census Rehearsal

Topic	Level	Purpose	Notes
Postcode	Household	Matching and analysis	
Address	Household	Matching	
Whether household was resident on census night	Household	Filter	Need to strictly apply census definition of usual residence – also identifies in movers
Tenure	Household	Analysis	
Type of Accommodation	Household	Matching	
Self-contained accommodation	Household	Matching	
Number of usual residents	Household	Quality assurance	Need to strictly apply census definition of usual residence
Response outcome	Household	Quality assurance	Non contact, refusal, vacant, second residence etc
Source of information	Household	Quality assurance	Householder, relative, neighbour, new resident, interviewer
Forename	Person	Matching	
Surname	Person	Matching	
Date of Birth	Person	Matching and analysis	
Estimated age	Person	Analysis	Used if no date of birth collected
Gender	Person	Analysis	
Simple marital status	Person	Matching and analysis	
Relationship to head of household	Person	For deriving household structure for analysis	
Full time student	Person	Filter	
Term time address	Person	Filter	
Simple ethnicity	Person	Analysis	Only broad classification suitable for analysis
Simple religion	Person	Analysis	This will be a known Northern Ireland variation – not required for England and Wales
Activity last week	Person	Analysis	
Migrant status (usual address 1 year ago)	Person	Analysis	Include a more expanded version for 2011
Country of birth – UK or non-UK	Person	Analysis	To identify internal and international migrants
Addresses and postcodes where household member could have been enumerated	Person	Measuring overcoverage	
Reason for other addresses	Person	Measuring overcoverage	New topic for 2011
Name of visitor on census night	Person	Matching	New topic for 2011
Date of birth of visitor on census night	Person	Matching and analysis	New topic for 2011
Gender of visitor on census night	Person	Analysis	New topic for 2011
Usual address and postcode of visitors on census night (or country)	Person	Matching	New topic for 2011
Intended length of stay	Person	Analysis	New topic for 2011. Required to obtain 12+ months usual residence population (i.e. to be able to filter out short term migrants)
Establishment type	Communal	Analysis	
Number of residents	Communal	Quality assurance	

- CCS fieldwork will start six weeks after Census Day. This is different from 2001, when the CCS commenced four weeks after Census Day. The timing of the fieldwork period is dictated by the need to wait until census fieldwork is finished (and thus maximise its response), balanced by the advantages of conducting the survey as soon as possible after Census Day
- Interviewing will be carried out in two stages: first, interviewers will identify every household within a postcode; second, they will then attempt to obtain an interview with a member of each household
- Unlike the census, identification of households within the interviewers' areas will not be guided by any list. Instead, maps of the CCS postcodes will be supplied to interviewers for them to confirm the physical extent of the postcodes on the ground by calling on households. To ensure interviewers visit every household in their allocated postcodes they will contact households adjacent but outside the postcode boundary to ensure that all households in the selected postcodes are included in the CCS. This process avoids the identification of households in the CCS being dependent on a potentially misleading address list
- To ensure the questionnaire will be short and simple, the CCS interview will ask for only a limited set of demographic and social characteristics for everyone living in a household, together with questions about the accommodation and simple relationship information. It will also ask probing questions about populations that are known to be missed, and also collect information on whether each resident could have been counted elsewhere. This is important, since we can only estimate for, or control, the adjustment for characteristics collected in the CCS. The topics that will be included in the CCS for the 2009 Census Rehearsal are listed in **Box two**
- To ensure census field staff do not make a special effort to obtain response in areas to be covered by the CCS, the CCS sample postcodes will be kept confidential and Census staff will be prevented from interviewing in the same area they had enumerated or managed
- Interviewers will be instructed to make as many calls as necessary to obtain an interview, and to call at different times and on different days to maximise the probability of making contact

Matching

Estimates of the total population will be based on a methodology known as dual system estimation. It is inevitable that some households and people will be missed by both the census and CCS but dual system estimation can be used to estimate this by considering the 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 it is necessary to match the records from the CCS with those from the census. It is essential that this matching process is accurate as the number of missed matches has a direct impact on the final population estimates. The 2011 matching strategy will be similar to that developed for the 2001 methodology by Baxter (1998), involving a combination of automated and clerical matching. The matching methodology and processes are currently undergoing a thorough review and, while there will be some improvements, the basic methodology and process outlined in **Box three** will remain unchanged.

Estimation of the population

Stage 1 – Dual system estimation

Dual system estimation (DSE), which was the approach used in 2001, is firstly used to estimate the population within the sample areas. The use of DSE requires a number of conditions to be met to ensure the minimisation of error in the estimates. These are fully discussed by Brown and Tromans (2007), but include:

- Independence between the census and CCS is required for an unbiased estimate. As a result the census and CCS will be operationally independent
- A closed population. It is assumed that households do not move in between the census and CCS. Clearly this will not be the case, and in 2011 this will be exacerbated by the longer time between the two
- Within an Output Area, the chance of a person being in the census or CCS is assumed to be the same across all people within the stratum (often called the homogeneity assumption). This is a reasonable assumption since Output Areas are small and contain similar types of people (Output Areas were designed to be internally homogenous with respect to the population)
- Perfect matching

After matching between the census and the CCS, a 2 × 2 table of counts of individuals or households can be derived. This is given in **Table 1**.

This output from the matching process will be used to estimate the undercount for each of the sampled Output Areas, using the data from the three postcodes sampled in each. Given the assumptions, DSE combines those people counted in the census and/or CCS and estimates those people missed by both by a simple formula to calculate the total population as follows:

$$\text{DSE} = n_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}}$$

Box three

The four key stages of the matching process

Stage 1 – Exact matching

CCS and census households and individuals where key details match exactly are automatically linked.

Stage 2 – Probability matching

CCS and census records that were not matched at Stage 1 of the process are then run through a probability matching process. A probability weight is assigned to each pair of CCS and census records based on the level of agreement between them. The higher the probability weight, the closer the agreement between the two records. Any household pairs with a high probability weight are linked and the individuals within them compared in a similar fashion.

Stage 3 – Clerical resolution

Pairs of households and individuals with a reasonable level of agreement are presented for clerical resolution. At this stage operators will simply be asked to determine whether the pair of records shown constitute a matching pair or not. They will not be expected to search for matching records.

Stage 4 – Clerical matching

The final stage of the matching process involves a clerical search for any census records corresponding to unmatched CCS households and individuals, using a set of strict matching protocols.

Table 1 2 × 2 Table of Counts of Individuals (or households)

		Census Coverage Survey		
		Counted	**Missed**	Total
Census	*Counted*	n_{11}	n_{10}	n_{1+}
	Missed	n_{01}	n_{00}	n_{0+}
	Total	n_{+1}	n_{+0}	n_{++}

This approach has been used widely for the estimation of wildlife populations (see Seber, 1982), and for estimating undercoverage in the US Census (see Hogan, 1993). The formula assumes that the proportion of CCS responders that were also counted in the census is identical to the proportion of CCS non-responders who were in the census (this is the independence assumption). Another explanation is that assuming independence, the odds of being counted in the CCS among those counted in the census should be equal to the odds of being counted in the CCS among those not counted in the census. The full derivation of the DSE is given by Brown (2000).

Research has shown that the application of the DSE at the Output Area level is relatively robust to small violations of the assumptions. However, serious violation of the assumptions can sometimes result in significantly biased estimates of the population. The lesson from 2001 is that there is likely to be some residual bias in the DSE due to failure of some of these assumptions. The section 'Adjustments to the population estimates' describes the proposed approach for making adjustments to the DSE to reduce any significant or substantial bias. In addition to making adjustments for bias, there will also be adjustments for the levels of estimated overcount.

The calculation of DSEs will be carried out for both individuals and households at Output Area level. The output from Stage 1 of the estimation process will be estimates of the true household and individual population for the CCS sampled postcodes.

Stage 2 – Estimation Area estimation

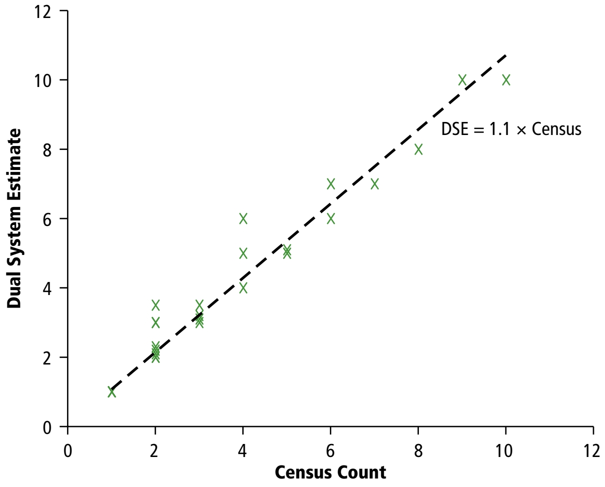
The second stage in the estimation process is to generalise the DSEs to the non-sampled areas.

As described in the CCS design section, LAs which do not have sufficient sample sizes to allow LA level estimates with an acceptable level of accuracy will be grouped together at the estimation stage into Estimation Areas. Within the Estimation Areas, a simple ratio estimator (which uses a straight line of best fit through the origin) will be used to estimate the relationship in the sample between the census count and the dual system estimate for each age-sex group within each HtC stratum, as shown in **Box four**. This relationship is then used to estimate the total Estimation Area population for each age-sex group in each HtC stratum by multiplying the census count by the estimated slope of the line. The variance of the estimate (a measure of accuracy used to construct confidence intervals) can also be estimated by standard methods that use replication techniques. The approach used in 2001 was a jackknife, which repeatedly calculates the estimate using a subset of the sample. Research is underway to see if alternative methods can provide better estimates of variance.

The output from this process will be estimates of the population for each Estimation Area by age and sex, together with an indication of their accuracy. A similar methodology will be used to calculate an estimate of the number of households, although this may use additional information.

Box four

The ratio estimator



Stage 3 – Local authority estimation

Since many Estimation Areas will consist of more than one LA, estimates of the age-sex population for each LA will need to be made. Small area estimation techniques (for a review of methods see Ghosh and Rao, 1994) can be applied to produce LA level population estimates that have lower variances (that is, smaller confidence intervals) than those using LA specific samples.

The small area estimation technique used will be similar to that employed in 2001. It uses information from the whole Estimation Area to model the undercount within the LAs, allowing for differences between them. This is where the Estimation Areas constructed of similar LAs will have additional benefit, as the small area model will not have to estimate large differences. The resulting population estimates will then be calibrated to the Estimation Area estimates, and their accuracy can also be calculated to provide confidence intervals around the LA population estimates.

Adjustments to the population estimates

In the 2001 Census, the quality assurance of the population estimates showed that there was some bias in the DSEs. As a result, Brown *et al* (2006) developed a method to make adjustments to the DSEs by incorporating additional external data. For 2011 the intention is to make corrections for any significant biases in the DSE as an integrated part of the methodology. However, some of these adjustments will not be possible until all the data have been processed. This section outlines three adjustments that are proposed for the DSEs – overcount, movers and residual dependence and correlation bias. The adjusted DSEs can be fed back into the usual ratio/small area estimation methods described above, so that the adjustments are then applied to the whole population and revised census estimates can be calculated. These adjustments fit nicely into the existing methodology and provide a mechanism for feeding in additional data.

Estimation of overcount

The 2001 methodology focused on measuring the population by adjusting for undercount. Overcount has not historically been a problem within UK censuses, and therefore measurement of it was given a low priority.

Based on its matching process, the England and Wales Longitudinal Study estimated that 0.38 per cent of the population responded twice. A study of duplicates within the census database confirmed this finding, estimating around 0.4 per cent (200,000) duplicate persons. However, no adjustments were made to the 2001 Census estimates for overcount.

One of the improvements to the coverage assessment methodology is a more rigorous measurement of overcount. Abbott and Brown (2007) presented a full discussion of the options for measuring overcount within the existing framework, concluding that a separate estimated adjustment at aggregate level should be made and that it should then be integrated into the DSE. They also recommended that a number of sources of information should be used to estimate the level of overcount.

The main type of overcount that can occur within the census is when an individual or household makes more than one return. An example of this is where a student is counted at their term-time address (correctly) and also counted at their home address (incorrectly) by their parents (where the parents fail to answer the term time address filter question correctly). This group, if not removed, would result in an overcount where they are incorrectly counted. This type of overcount is most associated with students, children of separated parents and people with a second residence.

In order to estimate this type of overcount, an automated matching process will be developed to search for duplicates in the census database, on a sample basis. The sampling strategy will use an approach where sampling continues until a pre-specified number of duplicates have been observed. The number to be observed is based upon the precision required for the estimation of the proportion of duplicates. The outcome will be estimates of duplication within the census by Government Office Region and broad demographic characteristics. These estimates will then be used to adjust the DSE estimates downwards.

The matching strategy to detect such duplicates efficiently is under development, but will be conservatively designed, to reduce the likelihood of false positive matches (that is, finding a duplicate when one does not exist). A clerical review of the possible duplicates will ensure the automated match is accurate. In addition, the England and Wales Longitudinal Study, which is a 1 per cent sample, will help to estimate the level of duplicates and provide a robust quality assurance. Lastly, information from the CCS will be used to estimate the geographical distribution, since we will not know which of the duplicates is correct (the CCS will define the correct location for duplicates within the CCS sample areas). Full details of the sampling and estimation strategies for duplicates are still being developed.

Movers

Households or individuals that relocate in the period between the census and CCS can cause a bias in the DSEs. If the coverage of movers is significantly lower than non-movers (a likely hypothesis, given that the census fieldwork process will find it hard to follow up movers), the DSE homogeneity assumption is violated, resulting in bias. To assess this we will use the CCS to collect information on movers that will allow an estimate of mover coverage, and make broad adjustments if that estimate is significantly lower than the estimate of coverage for the population.

Residual dependence and correlation biases

One or more of the assumptions that underpin the DSE will likely fail in some cases. Whilst the development of the DSE methodology has attempted to reduce the impact of assumption failures, there may be cases where there is a significant residual bias. This can only be detected by comparing the DSE results against alternative sources (which is the purpose of the quality assurance process shown in Figure 1). However, the source of the failure cannot be determined, and therefore any correction cannot be specific.

The methodology for correcting the DSE for bias requires a credible alternative population data source. The strategy for making an adjustment where a significant bias is detected is to develop the framework used in 2001, making it more realistic and including additional reliable sources of data. This will include the aggregate number of households in an area (from the census address register), census visitor data, demographic sex ratios, survey data or administrative sources. This piece of the methodology requires further development, and possible sources of data need further assessment of their quality. The possibility of using a third source at individual level and developing a triple system estimator has not yet been ruled out, but is very dependent on obtaining and matching high quality individual-level data sources.

Adjustment

Following the production of the census population estimates, the census database will be adjusted to take account of the undercount and overcount. The adjustment will be made on a 'net' basis – separate adjustments for undercount and overcount will not be made. Instead, the undercount adjustment will be reduced by the estimated level of overcount, and therefore (assuming that undercount is always larger than overcount) the adjustment will always be to add additional 'missed' records.

The estimated population defines the number of households and people to be imputed along with some basic information about coverage patterns for other characteristics. However, it is important to identify the detailed characteristics of those households and individuals missed by the census. The information on the characteristics of missed persons obtained in the CCS will be used to model the likelihood of households and persons, with their characteristics, being missed from the census. These models use the matched CCS/census data to predict (for example), the probability that a 20–24 year old male who is single, white, living in a privately rented house in the hardest to count stratum is counted in the census. It is crucial to note that the variables that are included in the models are those which are controlled explicitly by the adjustment process, and they have to be collected by the CCS.

Wholly missed households will be imputed, located using the census address register, and persons within counted households will also be imputed to account for those missed by the census. 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. This adjusted database will be used to generate all statistical output from 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. Tabulations derived from this database will automatically include compensation for these errors for all variables and all levels of geography, and will be consistent with the census estimated population.

Summary

The 2011 Census programme has a number of initiatives to improve the enumeration process and deliver a high quality product. This article outlines the proposed coverage assessment and adjustment methodology for the 2011 UK Censuses, and summarises the research carried out to date.

The proposed methodology meets the following key objectives of the coverage assessment strategy:

Key findings

- The 2011 Census coverage assessment methodology has been developed based on the 2001 methodology, taking into account the lessons learnt and the changes in the census design
- Improvements in the methods have been introduced following robust research using the information from 2001
- Innovations have been introduced, including the measurement of overcount, adjustments for bias in the DSE and more use of external data

(a) The methodology builds on the framework developed in 2001, with improvements designed to provide a more robust methodology or gains in precision for the key census population estimates. The key to this is the information from 2001, and this has led to some important improvements in the CCS design and estimation methodology. However, care has been taken to ensure the method is not optimised for the 2001 situation.

(b) Innovations include the development of methods for measuring overcount, and for detecting and adjusting residual biases in the DSE. These innovations recognise the changes in the census methodology and society, and are an important addition to the 2001 framework. However, it must be recognised that these do add complexity.

(c) To support the development of the methodology, stakeholders and users have been informed of progress throughout the development to allow input through many of the established consultation routes; this paper forms part of that process. Research papers have been published (see the reference list), and there is an ongoing series of documentation available through the ONS website. Easy to access documents have also been developed (see ONS, 2008) and there are plans to widen this further.

References

Abbott O (2007) '2011 UK Census Coverage Assessment and Adjustment Strategy'. *Population Trends*, 127, 7–14. Available at: www.statistics.gov.uk/downloads/theme_population/PopulationTrends127.pdf

Abbott O and Brown J (2007) Overcoverage in the 2011 UK Census. Paper presented to 13th Meeting of the National Statistics Methodology Advisory Committee. Available at: www.statistics.gov.uk/methods_quality/downloads/NSMAC13_Census_Overcoverage.pdf

Abbott O, Jones J and Pereira R (2005) '2001 Census Coverage Survey: Review and Evaluation', *Survey Methodology Bulletin,* 55, 37–47.

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 (2000) Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration. *University of Southampton*, unpublished PhD thesis.

Brown J, Abbott O, and Diamond I (2006) 'Dependence in the One-Number Census Project'. *J. R. Statist. Soc. A*, 169, 883–902.

Brown J, Diamond I, Chambers R, Buckner L and Teague A (1999) 'A Methodological Strategy for a One-Number Census in the UK'. *J. R. Statist. Soc. A*, 162, 247–267.

Brown J and Tromans N (2007) Methodological Options for Applying Dual System Estimation. Paper presented at ISI satellite conference. Available at: www.s3ri.soton.ac.uk/isi2007/papers/Paper22.pdf

Ghosh M and Rao J (1994) 'Small Area Estimation: An Appraisal'. *Statist. Sci.*, 9, 55–93.

Cabinet Office (2008) *Helping to shape tomorrow – The 2011 Census of Population and Housing in England and Wales*. Cm 7513. ISBN 9780101751322. The Stationary Office.

Hogan H (1993) 'The 1990 Post-Enumeration Survey: Operations and Results'. *J. Am. Statist. Ass.*, 88, 1047–1060.

Holt T, Diamond I, and Cruddas M (2001) 'Risk in Official Statistics: A Case-Study of the 2001 One-Number Census Project'. *J. R. Statist. Soc. D*, 50, 441–456.

Local Government Association (2003) The 2001 One Number Census and its quality assurance: a review. Research Briefing 6.03.

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

ONS (2008) *How are Census Estimates Produced?* Available at: www.ons.gov.uk/census/2011-census/process-info/statistical-meth/coverage-assessment-leaflet.pdf

Pereira R (2002) 'The Census Coverage Survey – The Key Element of a One Number Census'. *Population Trends*, 108, 16–23. Available at: www.statistics.gov.uk/downloads/theme_population/PT108.pdf

Seber G (1982) The estimation of animal abundance and related parameters. Second edition published by *Charles Griffin & Company Ltd*, London.

Statistics Commission (2004) *Census and Population Estimates and the 2001 Census in Westminster: Final Report*. Available at: www.statscom.org.uk/uploads/files/reports/census%202001.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, 495–522.