Advisory Group paper AG (08) 05 - 2011 UK Census Coverage Assessment and Adjustment Methodology

This paper outlines the proposed coverage assessment and adjustment methodology for the 2011 UK Census. This paper has been produced in order to inform users and stakeholders of the emerging methodology for measuring and adjusting for undercount and overcount, and to highlight where improvements over the 2001 Census methodology are being proposed. This paper is part of an ongoing process to consult with stakeholders about the methodology. Shorter, more accessible versions of the paper are available on request.

Action requested of advisory groups:

Advisory group members are invited to comment on the emerging methodological proposals either at the meetings or via correspondence by the end of June.

Owen Abbott
Census Methodology Division
Office for National Statistics
Segensworth Road
Titchfield
Fareham
Hants
PO15 5RR
Email: owen.abbott@ons.gov.uk

2011 UK Census Coverage Assessment and Adjustment Methodology

1. Introduction

1.1 The central objective of the 2011 Census is to provide high quality population statistics as required by key users such as policy makers and service providers, on a consistent and comparable basis for small areas and small population groups (ONS, 2004). The key mission critical aims include:

- provision of high quality, value-for-money statistics that meet user needs;
- maximising overall response rates and minimising non-response in specific areas and among particular population subgroups; and;
- building user confidence in the final results.

1.2 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 other sub-groups of the population such as age and sex groups. 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 population that are missed can be those which attract higher levels of funding. Therefore, without any adjustment, the allocations based upon the census would result in monies being wrongly allocated. It is therefore traditional that census undercount is measured and the outcome disseminated to users. Hence in order to achieve the mission critical aims outlined above, ONS outlined its coverage assessment and adjustment strategy in Abbott (2007). This paper outlines the proposed methodology for the 2011 UK Census arising from that strategy. Whilst the methodology is applicable to the UK, it is expected that there will be slight differences between countries to reflect local circumstances.

1.3 Section 2 provides background information on the methodology from previous UK censuses, and the lessons learnt from the most recent. The high level strategy is summarised in section 3, and then section 4 outlines the high level methodology. Sections 5 to 10 detail the methodological components and then after the plans for consultation are presented a summary of the paper is given.

2. Background

2.1 Most census taking countries undertake some form of coverage assessment and adjustment, usually using some form of post-enumeration survey (PES). Measured undercount levels have on the whole been increasing over the past few decades. More importantly, the differential nature of the undercount has worsened with, for example, young males in inner city areas becoming increasingly difficult to enumerate. This has led to increasing priority and focus on the methods for measuring this differential undercount.

The 2001 One Number Census

2.2 In the 2001 UK Census, the One Number Census (ONC) project had the goal of providing a methodology and processes to identify and adjust for the number of people and households not counted in the 2001 Census (see Brown *et al* 1999, Holt *et al* 2001). The aim was to provide a population estimate that would be the basis for the 2001 mid-year estimate (with a minor time lag correction), and for which all

census tabulations would add up to. The One Number Census measured the undercount in the 2001 Census to be 6.1 per cent of the total population.

2.3 The ONC 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. These are summarised by Chappell and Dobbs (2005) and ONS (2005).

Lessons Learnt

2.4 As a result, there were a number of key lessons from the ONC project that are pertinent to the strategy and methodology in 2011. These were explored by Abbott and Brown (2006). In summary, these lessons were:

- The ONC 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 CCS.
- Two of the weaknesses of the ONC 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

3. 2011 Coverage assessment and adjustment strategy

3.1 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 overriding strategy is to build on the ONC framework, using it as a platform to develop an improved methodology.

3.2 There are a number of other objectives:

- The strategy will address the lessons from 2001, looking for improvements and taking into account the changes to the census design.
- 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 be able to measure the extent of each of these, permitting more transparent adjustments.

- Local Authority District (in England and Wales) and age-sex level population estimates should aim for minimal variation of precision, therefore ideally being the same relative precision across all.
- Target precision rates (for sampling errors only) are 95% 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.
- Ensure that there are no Local Authorities 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 Local Authority total population that is wider than 6.1 per cent, and a 3 per cent confidence interval is the desirable upper bound).

4. Methodology

4.1 The methodology used to achieve the above 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 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. The CCS is covered in more detail in section 5.

b) The CCS records are matched with those from the Census using a combination of automated and clerical matching. The matching is covered in more detail in section 6.

c) The census database is searched for duplicates and the CCS is then used to help estimate the levels of overcount in the census. For more details see section 7.

d) The undercount is estimated within groups of similar Local Authorities (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. These DSEs are then used within some form of regression estimator to derive undercount estimates for the whole of the Estimation Area. The undercount estimation is covered in more detail in section 8.

e) The population estimates for the Estimation Areas are then calculated using the undercount and overcount estimates.

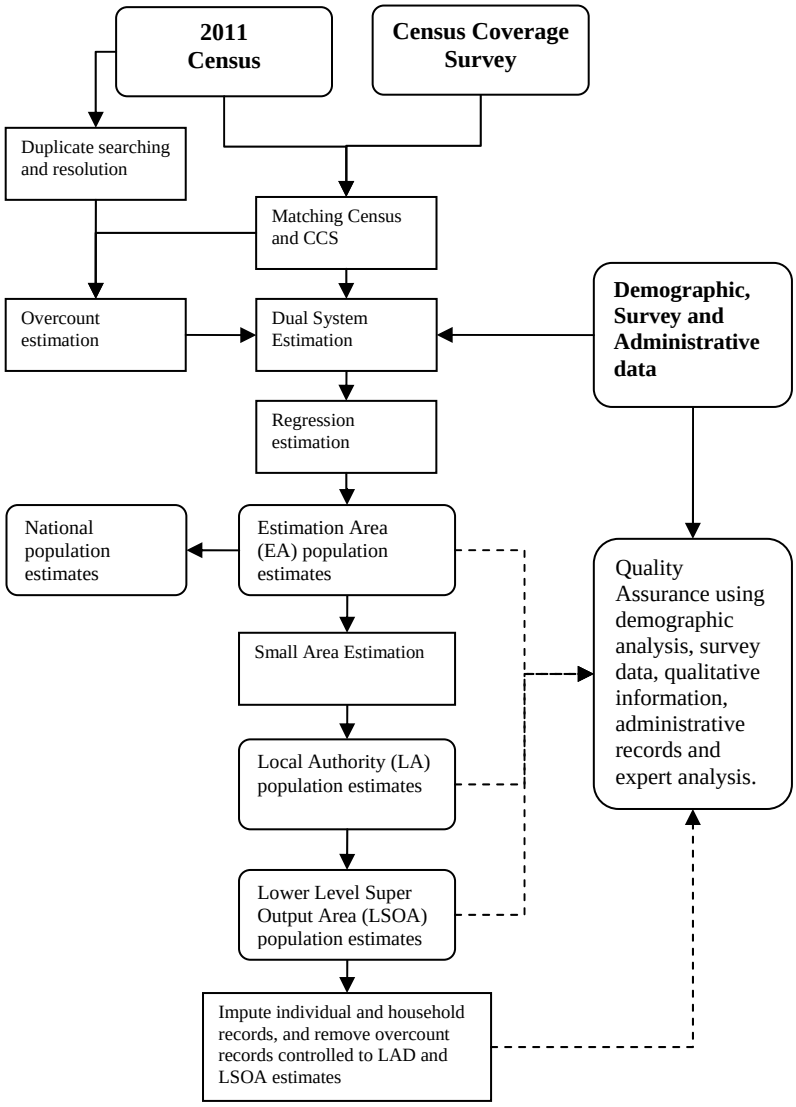
f) Small area estimation techniques will then be used to estimate the Local Authority (LA) population estimates. This is covered in section 8.

g) We are exploring the feasibility of using more complex small area estimation techniques, using additional data sources that will help to measure spatial variation, to estimate the population within the Lower layer Super Output Areas (LSOA) of each Local Authority. This will help control the adjustment process at a lower geographical level than LA and is covered in section 8.

h) Households and individuals estimated to have been missed from the Census will be imputed onto the Census database. In addition, a methodology will be developed to adjust for the households and individuals who are estimated to have been overcounted. These adjustments will be constrained precisely to the LA estimates but may not be constrained quite so precisely to the LSOA estimated populations. More detail can be found in section 9.

i) All the population estimates are quality assured using demographic analysis, survey data, qualitative information and administrative data to ensure the estimates are plausible. More detail can be found in section 10.

Figure 1 – The 2011 Coverage Assessment and Adjustment process overview



5. The Census Coverage Survey

5.1 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 around 16,500 postcode units across England and Wales. The sample of postcodes will be drawn from all Local Authority Districts (LADs).
- 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. This will be vital since the CCS, unlike the Census, is likely to be a voluntary survey.
- It will be operationally independent of the Census enumeration exercise.
- It will be undertaken during a four week period starting 6 weeks after Census Day.

Design

5.2 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, a national hard to count (HtC) index and by the size of key demographic subgroups. The second stage will then select 2 or 3 postcodes from within each selected Output Area. Output Areas replace the use of Enumeration Districts (EDs) in the previous CCS design, and we are selecting fewer postcodes in each primary sampling unit (allowing us to select more PSUs in the first place).

5.3 In 2001, the main geographic stratification in the design came from forming Estimation Areas (EAs) by grouping contiguous Local Authorities to create populations of around 500,000 persons and using these for sampling and estimation. However, for 2011, the strategy will be to draw the samples from Local Authorities directly, but then form the EAs at the estimation stage. This provides a sample that is better for making Local Authority level estimates – either directly for large LAs, or by using small area estimation for smaller LAs. A minimum sample size constraint within LAs will be applied. Where there is insufficient sample within an LA we will post-stratify the Local Authorities into Estimation Areas, grouping them by area type indicators (or perhaps census postback rates) rather than restricting the groups by geographical constraints. This is expected to increase the efficiency of the estimation process, as areas with similar undercount patterns will be grouped together. This can be done in advance to give an expected set of EAs which can then be confirmed or altered by a predefined strategy at the estimation stage.

5.4 As undercount is disproportionately distributed across areas, the OAs within each Local Authority are stratified according to a hard to count index. This index attempts to capture the variation associated with those characteristics most associated with undercount in a census. Goldring and Rahman (2007) use a modelling approach to identify the household characteristics most associated with undercount in the 2001 Census, and it is intended to update this research as new data on non-response becomes available prior to 2011. The top five variables identified at present (listed in order of importance) 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.

5.5 The index will be constructed from data that directly represents (or is highly correlated with) these factors, and where possible, is an up to date source. It is intended to construct the index using a similar methodology to that adopted for the Enumeration Targeting Categorisation developed for the 2007 Census Test by ONS (2006). This used a scaled ranking method to derive a score for Lower layer Super Output Areas (LSOAs), and then this is split into a national 60%, 20%, 10%, 8% and 2% categorisation. This is a more refined index than that used in 2001 (which had 3 levels and was based on information from a much smaller PES) because we have more confidence in the information about undercount patterns. In addition, the use of more up to date information should address one of the problems encountered in 2001 with the sample design in areas of high change.

5.6 The sample is then chosen from each of the hard to count strata within each Local Authority. Within each of the HtC categories the Output Areas will be further stratified by some kind of size strata, so that we select a sample that is approximately balanced across the key age-sex populations. The method of size stratification will aim to be simpler and more transparent than that used in 2001. Further work is required to define this, but it is expected that up-to-date small area population counts (or proxies) can be used to feed into this.

5.7 This sample design strategy should provide an efficient but robust design that spreads the sample across different area types. Finally, a number of postcodes (2 or 3) from each OA will be chosen at random. These selected postcodes will form the CCS sample.

5.8 It is likely that this sample design will be similar in Scotland and Northern Ireland, although there will be some differences due to geography and socio-demographic characteristics. For example, the 2001 CCS in Northern Ireland had a different definition for its hard to count index. It is expected that this will be the case again for the 2011 CCS Design.

Sample Size

5.9 The sample size of the CCS must be sufficiently large 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 and practicalities of carrying out a larger CCS. It is expected that a sample size similar to that employed in 2001 of around 16,500 postcodes or 320,000 households for England and Wales would provide an acceptable level of accuracy (relative confidence intervals of around 2%) for the populations of 500,000 (around 0.2% for the national population).

Survey Practicalities

5.10 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).

- CCS fieldwork will start six weeks after Census Day. This is a change 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 maximises 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 address within the postcode; second, they will then attempt to obtain an interview with a member of each household within the identified addresses.
- Unlike the Census, identification of addresses 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 addresses. 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 an 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 of everyone living in a household, 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 likely topics that will be included in the CCS are listed at Annex A.
- To ensure census staff will 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.

6. Matching

6.1 Estimates of the total population will be based on a methodology known as dual system estimation (see section 8). 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 number 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.

6.2 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.

6.3 The 2011 matching strategy will be similar to that developed for the 2001 ONC by Baxter (1998), involving a combination of automated and clerical matching. The matching process for a single CCS postcode is outlined below. There are four key stages:

Stage 1 - Exact Matching

6.4 CCS and Census households and individuals where key details match exactly are automatically linked. Households will only be considered matched at this stage when all individuals within the household pair have been linked.

Stage 2 - Probability Matching

6.5 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. For example, if a pair of records is identical with the exception of one detail, which may be due to recording error, then a high probability weight will be assigned. Any household pairs with a high probability weight are linked and the individuals within them compared. Only very similar households and individuals will be considered as matched at this stage.

Stage 3 - Clerical Resolution

6.6 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

6.7 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.

Quality Assurance

6.8 As previously mentioned, the accuracy of the matching process is critical to the accuracy of the population estimates. Quality Assurance procedures, similar to those used in the US, will be built into the matching process to ensure that the necessary high levels of accuracy are met. The output of the clerical matchers will be checked by expert matchers to ensure that all matched pairs of records are legitimate matches. These experts will also check that all unmatched records do not have a possible match using extensive database searches. A small number of supervisors will check the work of the expert matchers. These supervisors will also assist in marginal matching decisions. These processes should ensure accuracy and a consistent approach.

6.9 To estimate the overall accuracy of the matching, it is likely that a double matching strategy will be used. This is where the EA is independently matched twice, and results are compared. The level of discrepancies between the matching outcomes provides a measure of the accuracy of the clerical stage.

7. Measuring overcount

7.1 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. The 2001 CCS collected information about potential overcount by asking individuals whether there was anywhere else they might have been counted in the census. A matching study was undertaken based on the responses collected, resulting in an estimate of less than 0.1 per cent overcount. Further studies indicated that this might have been an under-

estimate. 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 backed up this finding, estimating that there was potentially around 0.4 per cent duplicate persons. These measures were probably underestimates, as they did not measure the numbers of people counted in the wrong location. However, no adjustments were made to the 2001 Census estimates for overcount.

7.2 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 adjustment at aggregate level should be made. They also recommended that a number of sources of information should be used to estimate the level of overcount. This section outlines the steps involved in measuring overcount and some of the sources of information that might be used. Further work is required to formulate how these are brought together to estimate the overall level of overcount.

Duplicate searches

7.3 One of the types 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 (correct) and also counted at their home address (*incorrect*) by their parents. This group, if not removed; 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.

7.4 In order to measure the level of this type of overcount, a matching process will be carried out to search for duplicates on the Census database. This could use a number of searching strategies to efficiently detect duplicates. Knowledge of the populations at risk of overcount can be used to help draw samples of those populations to carry out searches. For instance, one of the proposed new topics for inclusion in the 2011 Census is a question on second residence. This would ask where else in the UK each resident might be counted. This information could potentially help estimate global adjustments for duplicates via large-scale matching between census questionnaires.

7.5 Alternatively, another strategy is to look for name duplicates after blocking by date of birth across the entire census database. These strategies need further development work.

CCS information

7.6 Another type of overcount are people who are counted in the wrong place. An example is where a student is counted by their parents (*incorrect*), but missed where they should have been counted (their term time address). Nationally, these people in the adjusted data but once estimates are broken down by geography they become an overcount in one location and an undercount in the other. This type of overcount is difficult to detect and correct.

7.7 The CCS will collect information about where else residents might have been staying on Census night, as it did in 2001. However, in 2011 the CCS will also obtain information about visitors on census day and where those visitors could have been

counted. This information can be used, via matching, to establish adjustment proportions for those individuals who are counted in the wrong place.

Census Quality Survey

7.8 The last type of overcount are 'erroneous' enumerations. These are where the householder, enumerator or processing system:

a) creates fictitious people (e.g. pets);
b) includes people who are not usual residents of England and Wales (e.g. foreign visitors);
c) includes a baby born after census day; or;
d) includes someone who died before census day.

7.9 This group are a special problem as the only way they can be detected is by re-visiting householders and asking them to confirm that the people really exist and are usually resident there. The ONS is currently considering a survey to measure quality after the 2011 Census, which will involve re-interviewing households that provided a census return to measure the quality of response to all the questions. Whilst the sample size of the quality survey is unlikely to allow estimation at lower levels of geography (the sample size is currently planned to be around 2000 households), it could reveal if there has been an erroneous enumeration issue at national level. If this proves to be an issue, a national adjustment would then be required.

Longitudinal Studies

7.10 The Northern Ireland Longitudinal Study (NILS) is a 30% sample of the population that is updated by transactions from the NI health registration system. There is the possibility that it could be used to help measure overcoverage in Northern Ireland by:

a) detecting erroneous enumerations (by seeing whether unmatched census records could be found on NILS);
b) helping to resolve duplicates; and;
c) detecting and measuring individuals counted in the wrong location - again by looking for unmatched census records and finding if they are on NILS but in a different location.

8. Estimation of undercount

8.1 The next stage in the process is to derive estimates of the undercount for all Local Authority Districts (LADs) using the combined Census and CCS data generated by the matching. This section outlines the four stages in the process – the application of Dual System Estimation, the derivation of Estimation Area totals and the use of small area modelling to derive Local Authority totals and lower level estimates.

Stage 1 – Dual System Estimation

8.2 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.

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

		CCS		
		Counted	**Missed**	
Census	*Counted*	n_{11}	n_{10}	n_{1+}
	Missed	n_{01}	n_{00}	n_{0+}
		n_{+1}	n_{+0}	n_{++}

8.3 This output from the matching process will be used to estimate the undercount for each CCS postcode. This will be achieved using Dual System Estimation (DSE), which was the approach used in 2001. 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 Tromans and Brown (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.
- Within a postcode, 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 the majority of postcodes are small and contain similar types of people.
- Perfect Matching. This is the reason for requiring a high level of accuracy in the matching process described in section 6.

8.4 Given the assumptions, DSE combines those people counted in the Census and/or CCS and estimates those people missed by both by a relatively simple formulae to calculate the total population as shown below:

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

8.5 However, violation of the assumptions results in biased estimates of the population. In the 2001 ONC process, 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 data.

8.6 For the 2011 coverage assessment methodology, correcting for such biases in the DSE will be a part of the methodology. The strategy is to develop the framework used in 2001, making it more realistic and including additional reliable sources of data. This is likely to include the aggregate number of households in an area (perhaps from the Census household frame), census visitor data, demographic sex ratios, survey data or administrative sources. This piece of the methodology requires further development, and the sources of data that could be used also need further consideration. This is another area where there may be different sources used by Scotland or Northern Ireland. For instance, the NILS data may be able to be used to create local sex ratios, or even be used to adjust the DSEs by forming a Triple System Estimator.

8.7 The calculation of DSEs will be carried out for both individuals and households at postcode level. However, the household level DSEs may use additional information from the census household frame in the calculations in a different way to the bias adjustment strategy for individuals described in 8.6. The proposed methodology uses a modified triple system estimation approach, where the census address checking process provides the third list. However, the use of such an approach requires further research and a good understanding of the likely qualities of the census household frame before the final choice of methodology can be made.

8.8 The output from Stage 1 of the estimation process will be a set of estimates of the true household and individual population for the CCS sampled postcodes.

Stage 2 – Estimation Area estimation

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

8.10 As noted in the CCS design section, the Estimation Areas will be formed at the estimation stage. Therefore the first step is to confirm or alter the initial set of EAs through a predefined strategy. This will involve checking available data about the LAs (possibly data from the census field process such as postback rates) to see if there is evidence to suggest that any regrouping will be required.

8.11 Within the Estimation Areas, a form of regression estimator 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 Hard to Count stratum. This relationship is then used to estimate the total Estimation Area level undercount for each age-sex group in each HtC stratum. The variance of the estimate (which is a measure of quality) can also be calculated by a standard method called 'jacknifing', which repeatedly calculates the estimate using a subset of the sample.

8.12 The output from this process will be estimates of the undercount for each Estimation Area by age and sex, together with an indication of its accuracy. To obtain the total population estimate, the undercount estimate is added to the Census count which has been adjusted for the measured level of overcount (see section 7). A similar methodology will be used to calculate an estimate of the number of households, although this may use additional auxiliary information. All of the subsequent stages described below will be consistent with these population (and household) estimates.

Stage 3 – Local Authority District Estimation

8.13 Since many Estimation Areas will consist of more than one LAD, estimates of the age-sex (and household) population for each LAD will need to be made. This forms the third stage of the estimation process.

8.14 Many LADs, despite designing the CCS sample at this level, are unlikely to contain sufficient CCS postcodes to enable accurate direct estimates of population to be made. Small area estimation techniques can be applied to produce LAD level population estimates that have lower variances (i.e. smaller confidence intervals) than those that would be produced by just using the sample specific to each LAD.

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

Stage 4 – Lower Layer Super Output Area Estimation

8.16 A new stage being considered for the coverage assessment process is the calculation of population (and household) estimates for areas smaller than Local Authorities, ideally Lower Layer Super Output Areas (LSOAs). The rationale for doing this is to provide more accuracy in the adjustment process by providing lower level control totals. The methodology is yet to be developed but is likely to use more

complex small area estimation techniques than those used to derive the LA level estimates, since in general there will not be sample in all of the LSOAs. Therefore the methodology may make use of additional data sources (perhaps the census household frame, census postback rates or ONS small area population estimates) to increase the precision of the estimates. The estimates would be calibrated precisely to the LAD totals previously estimated, and confidence intervals would also be calculated – these will be used to ensure the final adjustment process (see section 9) gets as near as possible to the LSOA estimates but there might not be a guarantee that the final database will achieve them exactly.

9. Adjustment

9.1 Following the production of the population estimates at all levels, the census database will be fully adjusted to take account of the undercount and overcount. The exact approach has not yet been decided, particularly around how to adjust the database for the measured overcount. Therefore this section only describes the strategy for undercount adjustment.

9.2 The information on the characteristics of missed persons obtained in the CCS will allow the creation of a database which represents our best estimate of the entire population, whether counted by the Census or not. 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. 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.

9.3 The population estimates define the number of households and people to be imputed along with some basic information about coverage patterns for other characteristics. However, it is important that we identify the detailed characteristics of those households and individuals missed by the Census. The imputation process can be summarised in three stages.

Stage 1 – Modelling characteristics

9.4 The first stage of the process is 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) p, 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. These models have not yet been fully developed but they are likely to be fitted at regional level (to ensure sample sizes are adequate) and will be designed to provide the probabilities of different types of undercount – that is wholly missed households and persons missed from counted households. 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.

9.5 These predicted probabilities are then converted into coverage ‘weights’ (by taking the reciprocal of p). These weights will then be calibrated precisely to the population estimates at LAD level described in section 8, as these population estimates are the higher quality benchmark.

Stage 2 – Imputation of missed households and individuals

9.6 The second stage of the process will impute the wholly missed households and individuals (both within the wholly missed households and counted households), using the coverage weights to determine the characteristics of the imputations.

9.7 The weights are allocated to each Census household corresponding to the likelihood of households of that type being missed by the Census. The Census households are ordered by these weights and cumulative actual and weighted counts calculated. The cumulative counts are compared and, if the weighted count exceeds the unweighted count by more than 0.5, an imputed household is created with the characteristics of the current household. These characteristics will be limited to those used by the models and those which need to be controlled. Thus a number of 'skeleton' households are created that have certain characteristics. The strategy is to control those characteristics for the imputation process, but then use the Census item imputation system CANCEIS to complete the remaining information, since that will ensure the final data is consistent, preserving marginal distributions. In 2001 the system copied the whole household and individual records, and in some cases this resulted in the over-imputation of rare populations.

9.8 Alternatively, there is the possibility of using either late returned census data (ie questionnaires received after the CCS commences) or another individual level dataset (e.g. NILS) as ready made imputations to fill in some of the households prior to the use of CANCEIS. This would make use of real data, thereby potentially increasing the quality of the output database. However, this requires further development work before a decision is take on whether this is feasible.

9.9 Imputation of individuals missed from households counted by the Census is carried out in a similar fashion. The weights are used to impute individuals into the types of households that are likely to have missed people from their Census return.

Stage 3 – Placement of imputed households and individuals

9.10 Stage 3 involves the placement of the imputed households (and the individuals within them) and the placement of individuals into counted households.

9.11 The first of these will be achieved by using the information on the census household frame as a set of potential placement locations. The frame will include information about households that did not provide a return, but which the census enumerator indicated was occupied. The households to be imputed will be compared against the potential locations and scored based on their similarity to provide the best placement possible. This process will include the possibility of placing households in a 'new' address – that is one that is not on the household frame. These synthetic addresses will be allocated into a postcode to give them a geographical reference. The process will also try to ensure that the LSOA population estimates of households will be met (although we may have to use the confidence intervals around the LSOA level estimates as our constraints if we cannot guarantee to constrain precisely).

9.12 The individuals to be imputed into counted households will be placed into relevant household types (e.g. missed a baby from a 4 person household containing Mother, Father and young child). The relationship information will be modified to ensure consistency. This imputation process will ensure that the LSOA population estimates of persons should be met (as above), but also the LAD population estimates by age and sex are met exactly.

9.13 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 estimates.

10. Quality Assurance

10.1 A quality assurance process will be undertaken to ensure that the population (and household) estimates are sensible 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 being developed in 2008 but is likely to be similar to the model used in 2001 (described in White *et al*, 2006), albeit expanded to include more data sources and more comparisons. The critical part of this process is the selection of the data sources. Below is an example list (which is far from exhaustive) of those which are potential sources that could be used in the Quality Assurance process:

- Annual mid-year population estimates;
- Numbers of households paying council tax;
- Numbers of people listed on patient registers;
- Numbers of armed forces personnel;
- Numbers of children for whom child benefit is being paid;
- Numbers of people drawing the state pension (or an alternative benefit);
- Numbers of children at school;
- Numbers of students in higher education;
- Visitor data collected in the 2011 Census;
- Estimates of population characteristics from large surveys (e.g. the Integrated Household Survey);
- Information from Longitudinal Studies;
- Mortality Ratios; and;
- Sex ratios.

10.2 In addition, a range of descriptive information will be gathered to give a fuller picture of the area under consideration. This may include: demographic makeup of the areas considered; information about the conduct of the 2011 Census and the Census Coverage Survey; management information from the census processing operation; information on the estimation process; details of previous census coverage adjustments; and intelligence gathered on population estimates and the data sources.

10.3 A team of analysts will consider the evidence and summarise the key findings for each area. This information will be available to a panel consisting of specialist demographers, census managers and methodologists. ONS is also considering the possibility of involving independent user representatives on this panel. The panel will consider the evidence for each Estimation Area and LAD before either accepting or rejecting the estimates. In the event of any estimates being rejected at any stage of the process, a number of predefined adjustment and contingency strategies will be developed and be available to be used. This might include a strategy that uses a plausible target sex ratio to estimate the young male population, assuming the estimates of females are correct. Another strategy might be to re-estimate using

different post-strata (either Estimation Areas, the HtC index or individual or household characteristics).

10.4 The QA process will also include consideration of regional, national and special population estimates. The range of data may be different at that level, for example survey outputs will be suitable for comparing against population characteristics.

11. Consultation

11.1 To support the development of the methodology it is intended to keep stakeholders informed of progress and allow input through many of the established consultation routes, this paper being a part of that process. Some research papers have already been published (see the reference list), and there will be an ongoing series of such documentation, all made available through the ONS website.

11.2 A broad timetable outlining the key communication steps is given below (noting that the timetable is not fully agreed):

- Autumn 2008 – Methodology paper presented at stakeholder Workshops/Roadshows (TBC)
- Summer 2009 – Updated methodology papers with formal ONS on-line consultation
- Spring/Summer 2010 - Final Methodology paper(s) circulated and presented at a variety of forums
- Summer 2011 – Further communication in conjunction with other census consultations
- Autumn 2012 – Census results released, with associated coverage assessment and adjustment metadata

12. Summary

12.1 The 2011 Census project has a number of initiatives to improve the enumeration process and deliver a high quality census. This paper outlines the proposed coverage assessment and adjustment methodology for the 2011 UK Census. The development of this methodology for 2011 is underway, and this paper represents the research carried out to date and the intended direction of the methodological development.

12.2 Stakeholder management is also an important part of the strategy to ensure that key users both buy into and understand the methodology. This document is part of that, and it is intended that the methodology will be updated yearly to provide users an opportunity to provide ongoing feedback on the methodological proposals and developments. In addition, the research documentation will be made available on the National Statistics Census website for more technical users and higher level ‘easy to understand’ guides will be developed for users who do not wish to delve into the full details of the methodology.

References

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

Abbott, O. and Brown, J. (2006) A review of the 2001 One Number Census methodology and lessons learnt. Paper presented at GSS Methodology Conference, London, June 2006. Available at www.statistics.gov.uk/events/gss2006/downloads/D1Abbott.doc

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. Forthcoming.

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

Baxter (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**, 247-267.

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.J. and Tromans, N. (2007) Methodological Options for Applying Dual System Estimation. Internal ONS research paper. Available on request.

Chappell, R. and Dobbs, J. (2005) Are densely populated inner city areas easy to measure and estimate? Lessons learned from the 2001 Census in England and Wales. Paper presented at IAOS satellite meeting, Wellington, New Zealand, April 2005. Available at www.stats.govt.nz/about-us/events/satellite-meeting/default.htm

Holt, T., Diamond, I. D., 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, Available at www.lga.gov.uk/Documents/Publication/onenumbercensus.pdf

ONS (2004) 2011 Census: Strategic aims and key research in England and Wales. Information Paper. Available at www.statistics.gov.uk/downloads/theme_population/Strategic_aims.pdf

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

ONS (2006) Enumeration Targeting categorisation to be used in the 2007 Census test. Information Paper. Available at www.statistics.gov.uk/census/pdfs/EnumerationTargetingCategorisation.pdf

Rahman and Goldring (2007) Modelling Census household non-response. Paper presented at ISI Satellite conference, Southampton, August 2007. Available at www.s3ri.soton.ac.uk/isi2007/papers/Paper13.pdf

Statistics Commission (2003) The 2001 Census in Westminster: Final Report. Available at www.statscom.org.uk/media_html/reports/report_022/contents.asp

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.

White, N., Abbott, O., and Compton, G. (2006) Demographic analysis in the UK Census: a look back to 2001 and looking forward to 2011. 2006 Proceedings of the American Statistical Association, Survey Research Section [CD-ROM], American Statistical Association, Alexandria, VA.

Annex A – Likely Census Coverage Survey Topics

Ref	Topic	Level	Purpose	Notes
H1	Postcode	**Household**	Matching and analysis	
H2	Address	**Household**	Matching	
H3	Whether household was resident on census night	**Household**	Filter	Need to strictly apply census definition of residence
H4	Tenure	**Household**	Analysis	
H5	Type of Accommodation	**Household**	Matching	
H6	Self contained accommodation	**Household**	Matching	
H7	Number of usual residents	**Household**	Quality Assurance	Need to strictly apply census definition of residence
H8	Response outcome	**Household**	Quality Assurance	Non Contact, Refusal, Vacant, Second Residence etc
H9	Source of information	**Household**	Quality Assurance	Householder, Relative, Neighbour, New resident, Interviewer
P1	Forename	**Person**	Matching	
P2	Surname	**Person**	Matching	
P3	Date of Birth	**Person**	Matching and analysis	
P4	Estimated age	**Person**	Analysis	if no date of birth collected
P5	Gender	**Person**	Analysis	
P6	Simple Marital Status	**Person**	Analysis	
P7	Relationship to head of household	**Person**	For deriving household structure for analysis	
P8	Full time Student	**Person**	Filter	
P9	Term time address	**Person**	Filter	
P10	Simple Ethnicity	**Person**	Analysis	Only broad classification suitable for analysis

P11	Simple Religion	**Person**	Analysis	This will be an known Northern Ireland variation – not required for E&W
P12	Activity Last week	**Person**	Analysis	
P13	Migrant status (usual address 1 year ago)	**Person**	Analysis	Include a more expanded version for 2011
P14	Addresses and postcodes where HH member could have been enumerated	**Person**	Measuring overcoverage	
P15	Reason for other addresses	**Person**	Measuring overcoverage	New topic for 2011
P16	Time spent at other address	**Person**	Analysis	New topic for 2011. Used to model different population definitions
P17	Name of visitor on census night	**Person**	Matching	New topic for 2011
P18	Date of birth of visitor on census night	**Person**	Matching and analysis	New topic for 2011
P19	Gender of visitor on census night	**Person**	Analysis	New topic for 2011
P20	Usual address and postcode of visitors on census night (or country)	**Person**	Matching	New topic for 2011
P21	Intended length of stay	**Person**	Analysis	New topic for 2011. This might be required for modelling population staying less than 12 months.
C1	Establishment type	**Communal**	Analysis	
C2	Number of residents	**Communal**	Quality Assurance	