UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/AC.6/2008/3
29 February 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population and Housing Censuses

Eleventh Meeting
Geneva, 13-15 May 2008
Item 2 of the provisional agenda

CENSUS QUALITY ASSURANCE AND EVALUATION

2011 United Kingdom Census Coverage Assessment and Adjustment Methodology

Note by the Office for National Statistics, United Kingdom

Summary

This document outlines the proposed methodology for the 2011 UK Census. Section I presents the main objectives of the 2011 UK Census. Section II provides background information on the methodology from the 2001 UK Census, and the lessons learnt. The high level strategy is summarised in section III, and then section IV outlines the high level methodology. Sections V to X detail the methodological components and then a summary of the paper is given.

I. INTRODUCTION

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

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

2. There are a number of new innovations in the design of the 2011 Census to help meet these objectives. For the first time, a post-out methodology will be adopted, which will rely on the construction of a reliable household frame and a robust publicity campaign. This allows resources to be redirected into the follow-up operation, where the field force will be flexible in order to be able to react quickly to areas of poor response. Finally, respondents will be able to make their return over the internet.

3. Whilst these measures are designed to enable a high response, it is accepted that not everyone will be counted in the 2011 Census. 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. In the UK it is traditional that census undercount is measured and the outcome disseminated to users. 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.

4. Section II provides background information on the methodology from the 2001 UK Census, and the lessons learnt. The high level strategy is summarised in section III, and then section IV outlines the high level methodology. Sections V to X detail the methodological components and then a summary of the paper is given.

II. BACKGROUND

5. Most traditional census taking countries undertake some form of coverage assessment, usually using some form of post-enumeration survey (PES). Measured undercount levels have on the whole been increasing over the past few decades. The differential nature of the undercount is important since, for example, young males in inner city areas are difficult to enumerate. This has led to increasing priority and focus on the methods for measuring this differential undercount.

A. The 2001 One Number Census

6. In the 2001 UK Census, the One Number Census (ONC) project had the goal of providing a methodology to identify and adjust for the number of people and households not counted in the 2001 Census (see Brown et al 1999). The aim was to provide robust population estimates for the 376 Local Authority Districts (LADs - the key local government unit to which central funds are distributed) that would be the basis for the 2001 demographic mid-year estimate, 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. There were some issues with the results which led to further studies, adjustments and a number of lessons summarised by ONS (2005). In summary, these lessons were:

(a) the ONC was not able to make adjustments in all situations, particularly when there were pockets of poor census response;
(b) engagement with stakeholders is critical;
(c) that the methodology needs to be robust to failures in underlying assumptions and in particular have inbuilt adjustments for such failures;
(d) two of the weaknesses of the ONC were not having additional sources of data to complement the PES, and the perception that it would solve all ‘missing data’ problems;
(e) the measurement of over count requires greater attention and;
(f) the balance of 'measurement' resource between easier and harder areas needs careful consideration.

III. 2011 UK COVERAGE ASSESSMENT AND ADJUSTMENT STRATEGY

7. 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. Other objectives include:

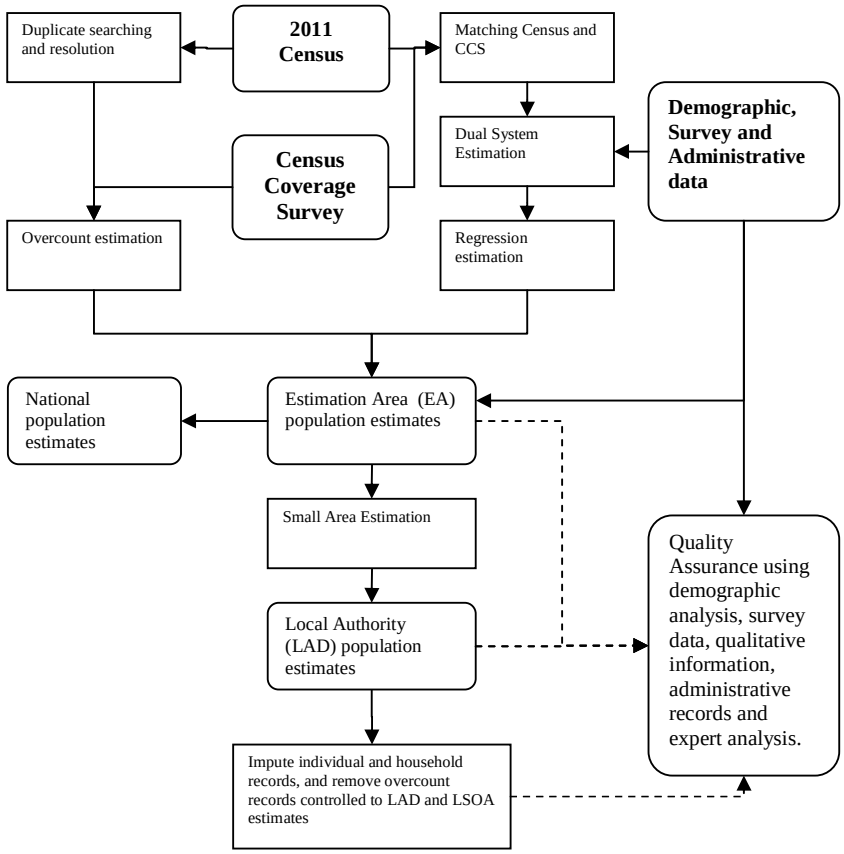
(a) the strategy will address the lessons from 2001, looking for improvements;
(b) gaining acceptance of the methodology from users. Users will not accept their census population estimates if they are not confident about the methodology used to derive them;
(c) target precision rates (for sampling errors only) are 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.

IV. METHODOLOGY

8. The methodology that is being developed to achieve the above strategic objectives is described in sections 5 to 10. This paper reports the current design of these methods, and the reader should be aware that as the development proceeds some aspects may be revised. 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 District;
(b) the CCS records are matched with those from the Census using a combination of automated and clerical matching;
(c) the census database is searched for duplicates and the CCS is then used to help estimate the levels of overcount in the census;
(d) the undercount is estimated within groups of similar Local Authority Districts (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 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;
(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 District population estimates;
(g) Households and individuals estimated to have been missed from the Census will be imputed onto the Census database. These adjustments will be constrained to the LAD estimates;
(h) All the population estimates are quality assured using demographic analysis, survey data, qualitative information and administrative data to ensure the estimates are plausible.

Figure 1. The 2011 Coverage Assessment and Adjustment process overview



V. THE CENSUS COVERAGE SURVEY

9. The key element in the coverage assessment and adjustment methodology is the CCS. The survey will comprise an intensive enumeration of a representative sample of around 320,000 households. It will be undertaken during a four week period starting 6 weeks after Census Day and will be operationally independent of the Census enumeration exercise. A short, paper-based interviewer-completed questionnaire will be used (as opposed to the Census self-completion questionnaire) designed to minimise the burden on the public. This will be vital since the CCS, unlike the Census, is a voluntary survey.

A. Design

10. The survey will be designed to enable census population counts to be adjusted for undercount at the national, local and small area level. The sample will be area based to enable coverage of households and individuals within households to be measured. A sample of postcodes (units used by the mail system) will be drawn from all Local Authority Districts. A two stage stratified sampling strategy will be used, with the main stratification being geography (Estimation Areas and Local Authority Districts) and a Hard to Count (HtC) index. This index attempts to capture the variation associated with those characteristics most associated with undercount in a census. The top five variables identified through modelling 2001 Census patterns (listed in order of importance) are households:

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

11. These variables will be combined to form a national index, probably with a 60 per cent, 20 per cent, 10 per cent, 8 per cent and 2 per cent categorisation. This sample design strategy should provide an efficient but robust design that spreads the sample across different area types.

B. Sample Size

12. 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.

C. Survey Practicalities

13. 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).

(a) CCS fieldwork will start six weeks after Census Day. This is a change from 2001, when the CCS commenced three and a half weeks after Census Day. The timing of the

fieldwork period is dictated by the need to wait until census fieldwork is finished (and maximises its response), balanced by the advantages of conducting the survey as soon as possible after Census Day.

(b) 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.

(c) 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. This process avoids the identification of households in the CCS being dependent on an address list.

(d) 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.

(e) 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.

(f) 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.

VI. MATCHING

14. Estimates of the total population will be based on a methodology known as dual system estimation (see section VIII). 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:

(a) both the Census and CCS;
(b) the Census but not the CCS; and
(c) the CCS but not the Census.

15. 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. False negative matches (where matches are missed) create a positive bias in the population estimates.

16. 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. Both exact and probability matching techniques are used. The matching strategy is designed to ensure that the false negative match rate is minimised.

VII. MEASURING OVERCOUNT

17. 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. A study of duplicates within the census database estimated that there was potentially around 0,4 per cent duplicate persons. However, no adjustments were made to the 2001 Census estimates for overcount.

18. 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 estimate should be made. They also recommended that a number of sources of information should be used to estimate the level of overcount, including searching the database for duplicates and using the CCS to detect individuals who are counted in the wrong location.

VIII. ESTIMATION OF UNDERCOUNT

19. The next stage is to estimate the undercount for all Local Authority Districts (LADs) using the combined Census and CCS data generated by the matching. There are three stages in the process.

Stage 1 – Dual System Estimation within sampled areas

20. After matching between the Census and the CCS, a 2×2 table of counts of individuals or households can be derived for each sampled postcode. 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_{++}

21. This output from the matching process will be used to estimate the undercount for each CCS postcode, stratified by age and sex. 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 include:

(a) independence between the Census and CCS is required for an unbiased estimate;
(b) 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;

(c) perfect matching. This is the reason for requiring a high level of accuracy in the matching process described in section VI.

22. 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:

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

23. 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. 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, such as demographic sex ratios or administrative sources.

Stage 2 – Estimation Area estimation

24. The second stage in the estimation process is to generalise the DSEs to the non-sampled areas. 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 using standard variance estimation techniques.

25. 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 VII). All of the subsequent stages will be consistent with these population estimates – they are the ‘best’ census based estimates of population.

Stage 3 – Local Authority District Estimation

26. Since many Estimation Areas will consist of more than one LAD, estimates of the age-sex population for each LAD will need to be made. Many of the 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.

27. 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 LAD census population estimates.

IX. ADJUSTMENT

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

Stage 1 – Modelling characteristics

29. 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) 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.

Stage 2 – Imputation of missed households and individuals

30. 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 probabilities to determine the characteristics of the imputations. The imputed households (and the individuals within them) will be located by using the information on the census household frame as a set of potential placement locations. 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).

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

X. QUALITY ASSURANCE

32. A quality assurance process will be undertaken to ensure that the population 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 likely to be similar to the model used in 2001 (described in ONS, 2005), albeit expanded to include more data sources and more comparisons. The critical part of this process is the selection of the data sources. The types of sources that could be used in the Quality Assurance process are:

(a) demographic mid-year population estimates;
(b) numbers of people listed on health registers;

(c) social security information;
(d) education information;
(e) estimates of population characteristics from large surveys;
(f) information from Longitudinal Studies; and
(g) demographic analyses (such as sex or mortality ratios).

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

34. A panel consisting of specialists will consider the evidence for each Local Authority District before either accepting or rejecting the estimates. In the event of any estimates being rejected, 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 young females are correct. 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.

XI. SUMMARY

35. The 2011 Census project has a number of initiatives to improve the enumeration process and deliver a high quality census. Despite these efforts, the 2011 Census will both miss people and also count them more than once. Evaluation of such coverage errors is critical, and the majority of traditional census takers use a Post-Enumeration Survey for this purpose. In the UK, ONS have developed a framework for measuring coverage of its censuses based on the success of the 2001 One Number Census and its Census Coverage Survey. For the 2011 Census, ONS are looking to use this framework as a platform to deliver high quality census population statistics. These improvements will help to ensure that users of the 2011 Census data are confident in the results.

REFERENCES

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

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

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

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

UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/AC.6/2008/3
29 February 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population and Housing Censuses

Eleventh Meeting
Geneva, 13-15 May 2008
Item 2 of the provisional agenda

CENSUS QUALITY ASSURANCE AND EVALUATION

2011 United Kingdom Census Coverage Assessment and Adjustment Methodology

Note by the Office for National Statistics, United Kingdom

Summary

This document outlines the proposed methodology for the 2011 UK Census. Section I presents the main objectives of the 2011 UK Census. Section II provides background information on the methodology from the 2001 UK Census, and the lessons learnt. The high level strategy is summarised in section III, and then section IV outlines the high level methodology. Sections V to X detail the methodological components and then a summary of the paper is given.

I. INTRODUCTION

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

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

2. There are a number of new innovations in the design of the 2011 Census to help meet these objectives. For the first time, a post-out methodology will be adopted, which will rely on the construction of a reliable household frame and a robust publicity campaign. This allows resources to be redirected into the follow-up operation, where the field force will be flexible in order to be able to react quickly to areas of poor response. Finally, respondents will be able to make their return over the internet.

3. Whilst these measures are designed to enable a high response, it is accepted that not everyone will be counted in the 2011 Census. 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. In the UK it is traditional that census undercount is measured and the outcome disseminated to users. 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.

4. Section II provides background information on the methodology from the 2001 UK Census, and the lessons learnt. The high level strategy is summarised in section III, and then section IV outlines the high level methodology. Sections V to X detail the methodological components and then a summary of the paper is given.

II. BACKGROUND

5. Most traditional census taking countries undertake some form of coverage assessment, usually using some form of post-enumeration survey (PES). Measured undercount levels have on the whole been increasing over the past few decades. The differential nature of the undercount is important since, for example, young males in inner city areas are difficult to enumerate. This has led to increasing priority and focus on the methods for measuring this differential undercount.

A. The 2001 One Number Census

6. In the 2001 UK Census, the One Number Census (ONC) project had the goal of providing a methodology to identify and adjust for the number of people and households not counted in the 2001 Census (see Brown et al 1999). The aim was to provide robust population estimates for the 376 Local Authority Districts (LADs - the key local government unit to which central funds are distributed) that would be the basis for the 2001 demographic mid-year estimate, 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. There were some issues with the results which led to further studies, adjustments and a number of lessons summarised by ONS (2005). In summary, these lessons were:

(a) the ONC was not able to make adjustments in all situations, particularly when there were pockets of poor census response;
(b) engagement with stakeholders is critical;
(c) that the methodology needs to be robust to failures in underlying assumptions and in particular have inbuilt adjustments for such failures;
(d) two of the weaknesses of the ONC were not having additional sources of data to complement the PES, and the perception that it would solve all ‘missing data’ problems;
(e) the measurement of over count requires greater attention and;
(f) the balance of 'measurement' resource between easier and harder areas needs careful consideration.

III. 2011 UK COVERAGE ASSESSMENT AND ADJUSTMENT STRATEGY

7. 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. Other objectives include:

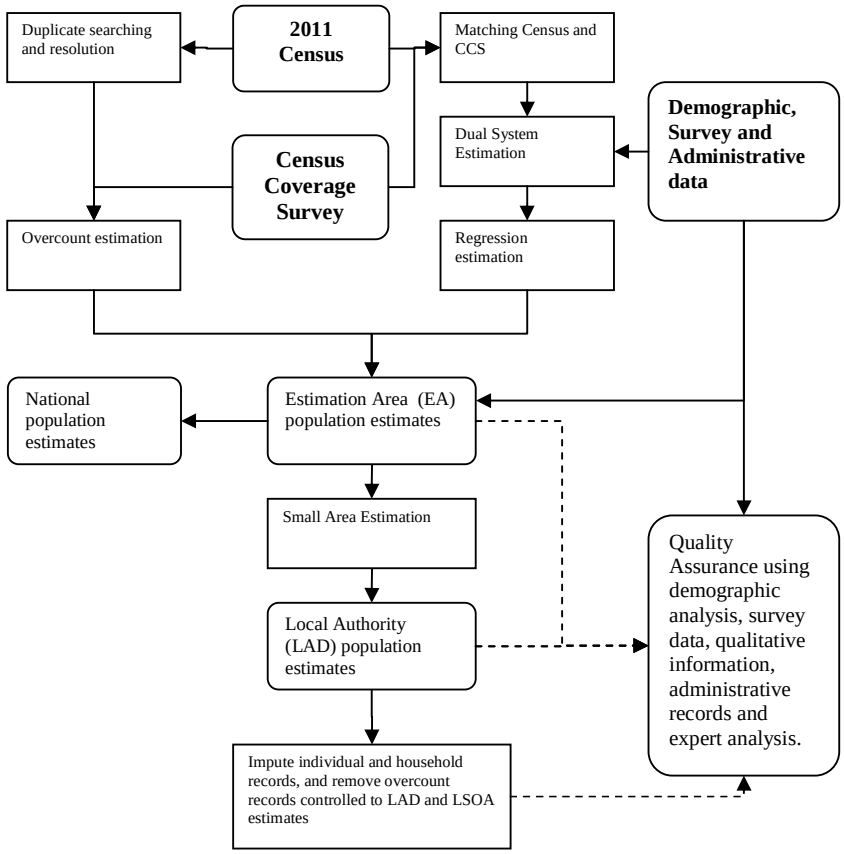
(a) the strategy will address the lessons from 2001, looking for improvements;
(b) gaining acceptance of the methodology from users. Users will not accept their census population estimates if they are not confident about the methodology used to derive them;
(c) target precision rates (for sampling errors only) are 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.

IV. METHODOLOGY

8. The methodology that is being developed to achieve the above strategic objectives is described in sections 5 to 10. This paper reports the current design of these methods, and the reader should be aware that as the development proceeds some aspects may be revised. 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 District;
(b) the CCS records are matched with those from the Census using a combination of automated and clerical matching;
(c) the census database is searched for duplicates and the CCS is then used to help estimate the levels of overcount in the census;
(d) the undercount is estimated within groups of similar Local Authority Districts (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 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;
(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 District population estimates;
(g) Households and individuals estimated to have been missed from the Census will be imputed onto the Census database. These adjustments will be constrained to the LAD estimates;
(h) All the population estimates are quality assured using demographic analysis, survey data, qualitative information and administrative data to ensure the estimates are plausible.

Figure 1. The 2011 Coverage Assessment and Adjustment process overview



V. THE CENSUS COVERAGE SURVEY

9. The key element in the coverage assessment and adjustment methodology is the CCS. The survey will comprise an intensive enumeration of a representative sample of around 320,000 households. It will be undertaken during a four week period starting 6 weeks after Census Day and will be operationally independent of the Census enumeration exercise. A short, paper-based interviewer-completed questionnaire will be used (as opposed to the Census self-completion questionnaire) designed to minimise the burden on the public. This will be vital since the CCS, unlike the Census, is a voluntary survey.

A. Design

10. The survey will be designed to enable census population counts to be adjusted for undercount at the national, local and small area level. The sample will be area based to enable coverage of households and individuals within households to be measured. A sample of postcodes (units used by the mail system) will be drawn from all Local Authority Districts. A two stage stratified sampling strategy will be used, with the main stratification being geography (Estimation Areas and Local Authority Districts) and a Hard to Count (HtC) index. This index attempts to capture the variation associated with those characteristics most associated with undercount in a census. The top five variables identified through modelling 2001 Census patterns (listed in order of importance) are households:

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

11. These variables will be combined to form a national index, probably with a 60 per cent, 20 per cent, 10 per cent, 8 per cent and 2 per cent categorisation. This sample design strategy should provide an efficient but robust design that spreads the sample across different area types.

B. Sample Size

12. 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.

C. Survey Practicalities

13. 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).

(a) CCS fieldwork will start six weeks after Census Day. This is a change from 2001, when the CCS commenced three and a half weeks after Census Day. The timing of the

fieldwork period is dictated by the need to wait until census fieldwork is finished (and maximises its response), balanced by the advantages of conducting the survey as soon as possible after Census Day.

(b) 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.

(c) 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. This process avoids the identification of households in the CCS being dependent on an address list.

(d) 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.

(e) 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.

(f) 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.

VI. MATCHING

14. Estimates of the total population will be based on a methodology known as dual system estimation (see section VIII). 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:

(a) both the Census and CCS;
(b) the Census but not the CCS; and
(c) the CCS but not the Census.

15. 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. False negative matches (where matches are missed) create a positive bias in the population estimates.

16. 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. Both exact and probability matching techniques are used. The matching strategy is designed to ensure that the false negative match rate is minimised.

VII. MEASURING OVERCOUNT

17. 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. A study of duplicates within the census database estimated that there was potentially around 0,4 per cent duplicate persons. However, no adjustments were made to the 2001 Census estimates for overcount.

18. 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 estimate should be made. They also recommended that a number of sources of information should be used to estimate the level of overcount, including searching the database for duplicates and using the CCS to detect individuals who are counted in the wrong location.

VIII. ESTIMATION OF UNDERCOUNT

19. The next stage is to estimate the undercount for all Local Authority Districts (LADs) using the combined Census and CCS data generated by the matching. There are three stages in the process.

Stage 1 – Dual System Estimation within sampled areas

20. After matching between the Census and the CCS, a 2×2 table of counts of individuals or households can be derived for each sampled postcode. 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_{++}

21. This output from the matching process will be used to estimate the undercount for each CCS postcode, stratified by age and sex. 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 include:

(a) independence between the Census and CCS is required for an unbiased estimate;
(b) 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;

(c) perfect matching. This is the reason for requiring a high level of accuracy in the matching process described in section VI.

22. 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:

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

23. 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. 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, such as demographic sex ratios or administrative sources.

Stage 2 – Estimation Area estimation

24. The second stage in the estimation process is to generalise the DSEs to the non-sampled areas. 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 using standard variance estimation techniques.

25. 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 VII). All of the subsequent stages will be consistent with these population estimates – they are the ‘best’ census based estimates of population.

Stage 3 – Local Authority District Estimation

26. Since many Estimation Areas will consist of more than one LAD, estimates of the age-sex population for each LAD will need to be made. Many of the 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.

27. 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 LAD census population estimates.

IX. ADJUSTMENT

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

Stage 1 – Modelling characteristics

29. 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) 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.

Stage 2 – Imputation of missed households and individuals

30. 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 probabilities to determine the characteristics of the imputations. The imputed households (and the individuals within them) will be located by using the information on the census household frame as a set of potential placement locations. 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).

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

X. QUALITY ASSURANCE

32. A quality assurance process will be undertaken to ensure that the population 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 likely to be similar to the model used in 2001 (described in ONS, 2005), albeit expanded to include more data sources and more comparisons. The critical part of this process is the selection of the data sources. The types of sources that could be used in the Quality Assurance process are:

(a) demographic mid-year population estimates;
(b) numbers of people listed on health registers;

(c) social security information;
(d) education information;
(e) estimates of population characteristics from large surveys;
(f) information from Longitudinal Studies; and
(g) demographic analyses (such as sex or mortality ratios).

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

34. A panel consisting of specialists will consider the evidence for each Local Authority District before either accepting or rejecting the estimates. In the event of any estimates being rejected, 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 young females are correct. 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.

XI. SUMMARY

35. The 2011 Census project has a number of initiatives to improve the enumeration process and deliver a high quality census. Despite these efforts, the 2011 Census will both miss people and also count them more than once. Evaluation of such coverage errors is critical, and the majority of traditional census takers use a Post-Enumeration Survey for this purpose. In the UK, ONS have developed a framework for measuring coverage of its censuses based on the success of the 2001 One Number Census and its Census Coverage Survey. For the 2011 Census, ONS are looking to use this framework as a platform to deliver high quality census population statistics. These improvements will help to ensure that users of the 2011 Census data are confident in the results.

REFERENCES

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

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

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

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