Method of Coverage Assessment and Adjustment in the UK 2011 Census of the Population

Taylor, Alan
Methodology Directorate,
Office for National Statistics,
Titchfield,
United Kingdom,
PO15 5RR
E-mail: alan.taylor@ons.gov.uk

Abbott, Owen; Sexton, Christine, Baillie, Mark
Methodology Directorate,
Office for National Statistics,
Titchfield,
United Kingdom,
PO15 5RR
E-mail owen.abbott@ons.gov.uk, christine.sexton@ons.gov.uk

Brown, James
Southampton Statistical Sciences Research Institute (S3RI),
University of Southampton,
Southampton,
United Kingdom,
SO17 1BJ, UK.
E-mail: jjb1@soton.ac.uk

ABSTRACT

The UK decennial population Census took place on March 27th 2011. Every effort was made to ensure that every member of the population was counted. However, no census is perfect and some people will have been missed. In terms of resource allocation, this can be an important issue as the people that are missed are often those who attract higher levels of funding. A smaller number will also have been counted twice, usually in two different locations. It is expected that this will be more of an issue in 2011 than previously due to nature of population changes. This Census builds on the successful methodology used in the 2001 Census developed to address the issue of undercount. The 2001 Census was the first UK Census to have a fully integrated coverage assessment and adjustment process. The aim was to provide population estimates that took account of undercount estimated using the census coverage survey through a dual system estimation methodology. In this talk I will describe the components of the 2011 methodology and how they have been development from those used in 2001. These components include the design of the coverage survey, use of dual system estimation techniques and how this is integrated to estimate population totals by administrative area. In addition, supporting evidence from simulations carried out by the ONS Census Coverage Team will be discussed.

Introduction

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

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

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

The 2001 methodology was a big step forward. However, there were a number of key lessons from the ONC project which were fully evaluated by ONS (2005). In summary, these lessons were:

- The methodology was not able to make adjustments in all situations, particularly when there were pockets of poor census response
- Engagement with stakeholders is critical
- 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
- The measurement of overcount requires greater attention
- The balance of 'measurement' resource between easier and harder areas needs careful consideration, as more sample in harder areas may even out the quality of the estimates

This presentation provides an overview of methodological framework used to assess and adjust for coverage in the 2011 Census. The detailed methodology for each of the components is summarized, including the design of the coverage survey, the estimation and adjustment process and the improvements that have been introduced. The presentation will highlight the results of simulation based research that has been used to validate methodological innovations. This presentation focuses on the methodology as applied to England & Wales. An equivalent methodology is used in Scotland and Northern Ireland.

2011 Coverage assessment and adjustment methodology

The overall methodology used in coverage assessment and adjustment is summarized in the following. The key stages are shown in Figure 1

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

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

c) A large sample of census records are checked to see if they are duplicates. The CCS is then used to help estimate the levels of overcount in the census, by broad age-sex groups and region.

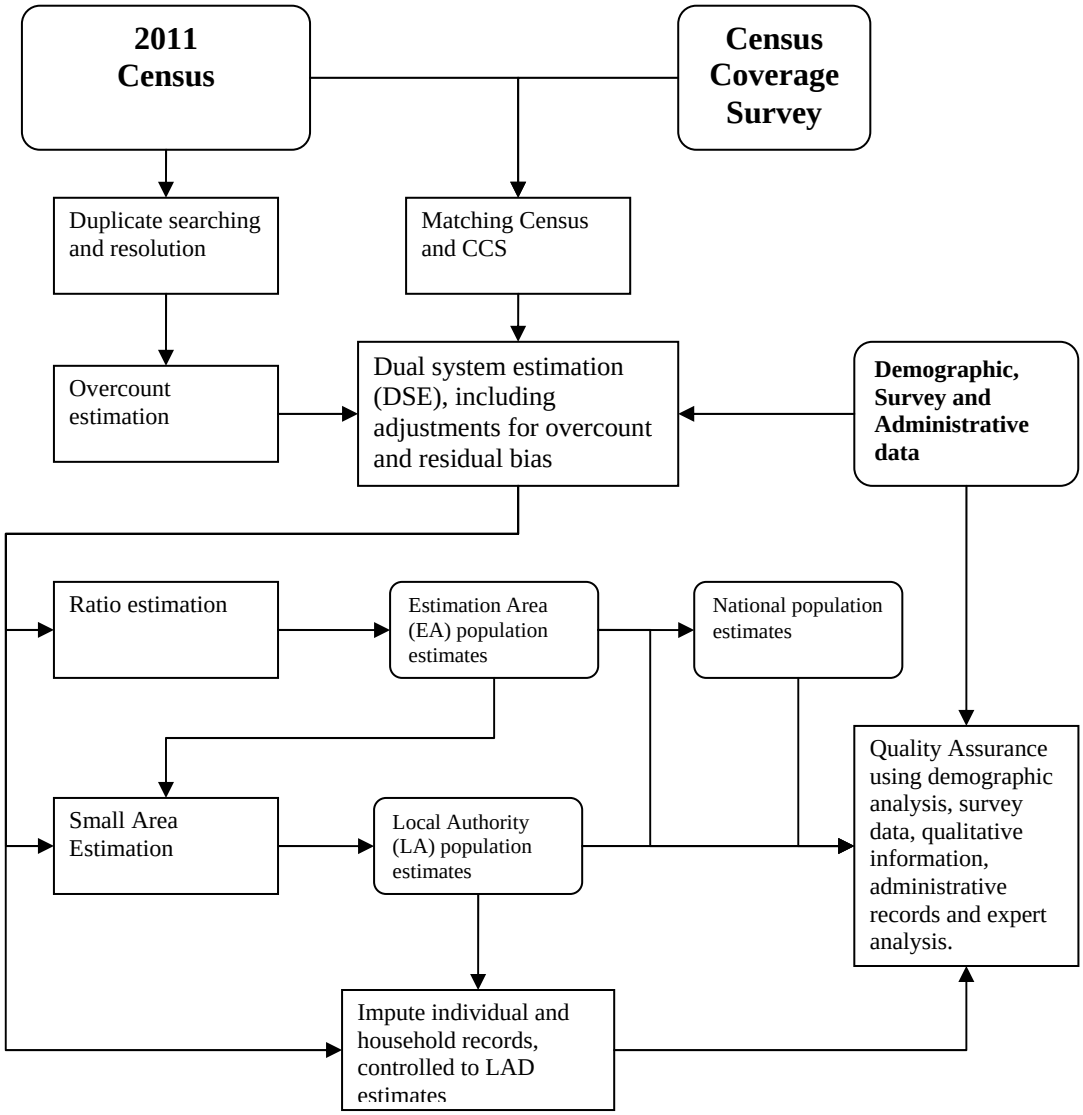
d) The population is estimated within groups of contiguous LAs (called Estimation Areas (EAs)) to ensure that sample sizes are adequate. The matched Census and CCS data are used within a dual system estimator (DSE), which is augmented with other reliable sources of data to estimate and adjust for any residual bias. The overcount propensities are used to adjust the DSEs so that a net adjustment is formed. These DSEs are then used within a simple ratio estimator to derive estimates for the whole of the Estimation Area.

e) Small area estimation techniques are used to estimate the LA population estimates.

f) Households and individuals estimated to have been missed from the Census are imputed onto the Census database. These adjustments are constrained to the LA estimates.

g) All the 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



The Census Coverage Survey

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

- It is designed to enable census population counts to be adjusted for underenumeration at the national, local and small area level
- It comprises of an intensive enumeration of a representative sample of postcodes. The sample of postcodes are drawn from all Local Authorities. The national sample size is approximately the same as was used in 2001
- It consists of a short, paper-based interviewer-completed questionnaire (as opposed to the Census self-completion questionnaire) designed to minimise the burden on the public, and therefore maximise response rates. This will be vital since the CCS, unlike the Census, will be a voluntary survey
- It is operationally independent of the Census enumeration exercise
- It is undertaken during a four week period starting six weeks after Census Day

Design

The CCS is a stratified two-stage sample selection of postcodes that will be independently re-enumerated. The first stage selected a sample of Output Areas (OAs) as the primary sampling unit, stratified by Local Authority and a national hard to count (HtC) index. Output areas are a standard statistical geography in England and Wales consisting of a relatively homogeneous area of approximately 125 households. The second stage then selected approximately half of the postcodes within the OAs to obtain a target of approximately 60 addresses within each selected OA. In 2001, five postcodes were selected in each primary sampling unit. In 2011 the clustering in the sample is reduced compared to 2001 making it more statistically efficient, but increasing travel costs slightly.

In 2001, the main geography in the design came from forming Estimation Areas (EAs) by grouping contiguous Local Authorities to create populations of around 500,000 people, and using these for sampling and estimation. However, for 2011 the strategy is to draw the samples from LAs directly and then to form the EAs at the estimation stage. This provides a sample that is better for making LA level estimates – either directly for large LAs, or by using small area estimation for smaller LAs. Where there is insufficient sample within an LA to estimate the population with an acceptable level of accuracy, LAs are grouped into Estimation Areas, based on geographical proximity. Simulation research demonstrated that there was no gain in precision by grouping LAs by area type indicators. This is because estimation is stratified by hard-to-count and age-sex strata which are the primary determinants of undercount.

As undercount is disproportionately distributed across areas, the OAs within each LA are stratified according to a national Hard-to-Count index. This index captures non-geographical variation in undercount. Research into the household characteristics most associated with undercount in the 2001 Census was used as the basis for the index. The underlying model that has been developed to predict the relative difficulty of enumerating an Output Area attempts to include up to date data sources. For example, the proportion of persons claiming unemployment benefits, a measure of the proportion of persons who are non-'White British' and a measure of the relative house price within an LA.

The use of more up to date information ensures that the sample design is robust in areas of high change. The national hard to count index partitions all OAs in England and Wales into a 40%, 40%, 10%, 8% and 2% distribution, which is similar to that used in 2001, but is more refined (the 2001 index had three levels with a 40%, 40%, 20% distribution) because we have more confidence in the information about undercount patterns.

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

Sample Size

The sample size of the CCS must be sufficiently large to ensure that the accuracy of the population estimates is acceptable. The larger the sample size, the more accurate the population estimates; however, this must be balanced against the cost, quality and practicalities of carrying out a larger CCS. Work has been undertaken to explore the precision of the estimates for different CCS sample sizes and Census coverage patterns. Based on this, a sample size similar to that employed in 2001 of around 17,000 postcodes (about 1.2%) or 325,000 households for England and Wales will provide an acceptable level of accuracy (relative confidence intervals of around 2%) for populations of 500,000 (around 0.2% for the national population).

Survey Practicalities

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

- CCS fieldwork started six weeks after Census Day. This is different from 2001, when the CCS commenced four weeks after Census Day. The timing of the fieldwork period is dictated by the need to wait until census fieldwork is finished (and thus maximise its response), balanced by the advantages of conducting the survey as soon as possible after Census Day
- Unlike the Census, identification of households within the interviewers' areas was not guided by any list. Instead, maps of the CCS postcodes were supplied to interviewers for them to confirm the physical extent of the postcodes on the ground by calling on households, testing the postcode boundary where necessary. This process avoids the identification of households in the CCS being dependent on a potentially misleading address list
- To ensure the questionnaire was short and simple, the CCS interview ask only a limited question set for everyone living in a household. It also asked probing questions about populations that are known to be missed, and also collected information on whether each resident could have been counted elsewhere.

Matching

Estimates of the total population are based on a methodology known as dual system estimation methodology. It is inevitable that some households and people will be missed by both the Census and CCS but dual system estimation can be used to estimate this by considering the numbers of the people observed by:

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

In order to identify the numbers in each of these groups it is necessary to match the records from the CCS with those from the Census. It is essential that this matching process is accurate as the number of missed matches has a direct impact on the final population estimates. The 2011 matching methodology was similar to that developed for the 2001 methodology by Baxter (1998), using a process of exact and

probabilistic matching with clerical review

Estimation of the population

Stage 1 – Dual System Estimation

Dual System Estimation (DSE), which was the approach used in 2001, is firstly used to estimate the population within the sample areas. The use of DSE requires a number of conditions to be met to ensure the minimisation of error in the estimates.

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

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

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

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

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

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

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

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

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

Stage 2 – Estimation Area estimation

The second stage in the estimation process is to generalise the DSEs to the non-sampled areas. As described in the CCS design section, LAs which do not have sufficient sample sizes to allow LA level estimates with an acceptable level of accuracy were grouped together at the estimation stage into Estimation Areas to ensure sample sizes are sufficient. Within the Estimation Areas, a simple ratio estimator (which uses a straight line of best fit through the origin) is used to estimate the relationship in the sample between the census count and the dual system estimate for each age-sex group within each HtC stratum. This relationship is then used to estimate the total Estimation Area population for each age-sex group in each HtC stratum by multiplying the census count by the estimated slope of the line. The variance of the estimate (a measure of accuracy used to construct confidence intervals) can also be estimated efficiently using standard bootstrap techniques. Research on the construction of EAs and the estimation of variance and confidence intervals at that level will be discussed in the presentation.

Stage 3 – Local Authority Estimation

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

The small area estimation technique used is similar to that employed in 2001. The method uses information from the whole Estimation Area to model the undercount within the LAs. Extensive simulation work based on 2001 data which compared the efficiency of a number of differing small area models found that a simple synthetic model provided the optimal results for the majority of Estimation Areas and their constituent LAs. Only where there was very marked undercount differentials across the constituent LAs within an EA did a model with additional LA specific effect prove more effective. This is primarily due to the stratification by Hard to Count and age-sex which remove the vast majority of LA specific undercount differentials. The method employed therefore applies the synthetic model. This model is replaced with a model with LA specific effects only when these are found to be significant. This research is discussed in an associated ISI presentation by Denise Silva.

Adjustments to the Population estimates

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

Estimation of overcount

The 2001 methodology focused on measuring the population by adjusting for undercount. Overcount

has not historically been a problem within UK censuses, and therefore measurement of it was given a low priority. Based on its matching process, the England and Wales Longitudinal Study estimated that 0.38 per cent of the population responded twice. A study of duplicates within the census database confirmed this finding, estimating around 0.4 per cent (200,000) duplicate persons. No adjustments were made to the 2001 Census estimates for overcount.

One of the improvements to the coverage assessment methodology is a more rigorous measurement of overcount. A separate estimated adjustment at aggregate level is integrated into the DSE. The main type of overcount that can occur within the Census is when an individual or household makes more than one return. This group, if not removed, would result in an overcount where they are incorrectly counted. This type of overcount is most associated with students, children of separated parents and people with a second residence.

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

The matching strategy is conservatively designed, to reduce the likelihood of false positive matches (i.e. finding a duplicate when one does not exist). A clerical review of the possible duplicates ensures the automated match is accurate. In addition, the England and Wales Longitudinal Study, which is a 1% sample, will provide an independent high quality estimate of the level of duplicates to validate the overcount estimate at the national level. Lastly, information from the CCS can be used to estimate the geographical distribution, since we do not know which of the duplicates is correct (the CCS will define the correct location for duplicates within the CCS sample areas).

Residual dependence and correlation biases

One or more of the assumptions that underpin the DSE will likely fail in some cases. Whilst the development of the DSE methodology and the implementation of the CCS aims to minimize the impact of assumption failures, there may be cases where there is significant residual bias. The two key assumptions behind the dual-system approach are homogeneity and independence. Failure of either of these assumptions leads to bias and while the failures are different, the outcome is equivalent, namely, an odds ratio not equal to one. Therefore, the overall adjustment strategy is based around an estimated odds ratio and has the potential to address failures of both assumptions.

Residual bias can only be detected by comparing the DSE results against alternative sources (which is the purpose of the Quality Assurance process shown in Figure 1). The methodology for correcting the DSE for bias is also based on using an alternative population count of the number of households. A number of methods for estimating the odds ratios underlying the DSE to adjust for dependence using this independent count were investigated. These build on the approach used in 2001 and are based on alternative methods of estimation of the household and person level odds ratio for use in the bias adjustment. The findings of the evaluation of these approaches will be reviewed in this presentation.

Adjustment

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

is always to add additional 'missed' records.

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

Wholly missed households are imputed, located using information on the location of missed households in the Census, and persons within counted households will also be imputed to account for those missed by the Census. This uses a similar methodology to that used in 2001, described by Steele et al (2002). The methodology has been refined to be more efficient and to remove the need for incremental adjustment to the imputation set due to a lack of fine control on the characteristics of the imputed individuals. This adjustment process imputes a subset of core characteristics for each individual as 'skeleton' records. These are then fully imputed using the Statistics Canada donor imputation program CANCEIS which more appropriately controls imputation across all Census variables. The research underlying these improvements will be discussed in the presentation.

The result is an individual level database that represents the best estimate of what would have been collected had the 2011 Census not been subject to undercount or overcount. Tabulations derived from this database will automatically include compensation for these errors for all variables and all levels of geography, and will be consistent with the census estimated population.

Summary

This presentation outlines the coverage assessment and adjustment methodology for the 2011 UK Census, and summarizes some of the research carried out in order improve the methodology from that used in 2001.

The proposed methodology meets the following key objectives of the coverage assessment process for the 2011 Census:

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

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

c) To support the development of the methodology, stakeholders and users have been informed of progress throughout the development to allow input through many of the established consultation routes. Research papers have been published and there is an ongoing series of documentation available through the

ONS website. As part of this process a full Independent Review of the Coverage Assessment and Adjustment process was carried out (ONS (2010))

REFERENCES

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

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

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

Baxter (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., Abbott, O., and Diamond I. (2006) Dependence in the one-number census project. *J. R. Statist. Soc. A*, **169**, 883-902.

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.

Ghosh, M. and Rao, J. N. K. (1994) Small area estimation: an appraisal. *Statist. Sci*., **9**, 55-93.

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

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

Holt, T., Diamond, I. 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.

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

ONS (2010) Independent Review Coverage Assessment, Adjustment and Quality Assurance. Available at www.ons.gov.uk/census/2011-census/2011-census-project/independent-assessments/independent-review-of-coverage-assessment--adjustment-and-quality-assurance/index.html

Pereira, R. (2007) The Census Coverage Survey – the key element of a One Number Census. *Population Trends*, **108**, 16-23. Available at www.statistics.gov.uk/downloads/theme_population/PT108.pdf

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.