The design of the 2011 Census Coverage Survey
Owen Abbott and Miguel Marques dos Santos, ONS

1. Introduction

Most census taking countries undertake some form of coverage assessment and adjustment, usually using some form of post-enumeration survey (PES). For the 2011 UK Census, the Office for National Statistics has committed to measuring coverage using a large Post-enumeration Survey, the Census Coverage Survey (CCS), as the key component of the overall coverage assessment and adjustment strategy (ONS, 2006a). The aim is to use the 2001 One Number Census as a platform to develop an improved coverage assessment and adjustment methodology.

The most important aspect of any survey is its design. No amount of statistical analysis can compensate for a badly designed survey. The main objectives for the CCS sample design are to select a robust sample, one that avoids bias and other sampling errors and achieves the maximum precision considering the available resources. This paper reports on the work that has been undertaken to develop the sample design for the 2011 CCS, and future work in this area.

The paper firstly outlines the design for the 2001 CCS which was used to measure coverage of the 2001 Census. It then moves on to the main aspects of the 2011 CCS design, focusing on where improvements are being sought over the approach used in 2001. This includes the choice of sampling units, geographic stratification, demographic stratification, sample size and sample allocation. The findings of studies conducted to date are discussed and the paper then outlines the further work that is planned for the design process.

2. The 2001 CCS Design

The 2001 One Number Census 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). The extent of the underenumeration was identified using a large survey covering approximately 320,000 households, the CCS, which provided an independent enumeration of a sample of areas. The aim of the CCS following the 2001 Census was to facilitate the estimation of underenumeration at a sub-national level (by age and sex), and to allocate this underenumeration down to small areas. The CCS was a stratified two stage sample, designed to be robust so that it could measure coverage across all groups (Brown *et al*, 1999). England and Wales were firstly split into 101 Estimation Areas (EAs), which consisted of contiguous groups of Local Authorities (LAs) with a combined population of about 500,000 persons. These formed the primary stratification, and samples were drawn from each of these strata. Within each EA, 1991 Census Enumeration Districts (EDs) were stratified firstly by a Hard to Count index (HtC) and secondly by size of key age-sex groups. The index was constructed from the 1991 Census information and distinguishes between enumeration districts based on their expected level of census coverage. It consisted of 3 strata - 'easy' (the 40% of 1991 EDs that had the lowest expected undercoverage), 'medium' (the next 40% of EDs) and 'hard' (the top 20% of EDs). The size strata were formed using a design variable that captured the age-sex structure of the 1991 EDs using babies, young males and elderly female age-sex groups. A sample of EDs (Primary Sampling Units (PSUs)) was chosen within these strata, and then a second stage selection chose a number of postcodes within each selected PSU. The number of postcodes chosen was dependent on the mean number of addresses per postcode within the selected ED.

The 2001 CCS was a success. Generally the design strategy worked well, and the CCS was able to provide the data to make robust adjustments across most (95% or more) areas and population subgroups. However, a number of studies of the estimation methodology and results highlighted that there were some problems related to the sample design. These problems were generally associated with sample balance. For instance, ratio estimation is more robust when the sample reflects the population distribution of the auxiliary (in this case the census) count. A number of modifications were made to the ratio estimator to make it more robust to a lack of balance, but it is perhaps better to achieve balance through the sample design rather than making modifications at the analysis stage (Abbott and Brown, 2006).

In addition, one of the weaker areas of the 2001 CCS was its reliance on out of date 1991 Census information used to construct the HtC index and size strata. There were some issues in areas that had undergone substantial redevelopment since the 1991 Census. In such areas, the HtC stratification was a poor predictor of undercoverage and so there was a greater chance that the sample did not include the hardest areas. This occurred in Manchester, where additional studies (see ONS, 2004) were required to provide a more robust population estimate. Lastly, the balance of 'measurement' resource between easier and harder areas needs careful consideration - for example how should the sample be allocated given a fixed overall size? Given that we have more information about undercount patterns than prior to the 2001 Census, this knowledge could be used to help increase the efficiency of the 2011 CCS Design.

3. Objectives of the 2011 CCS Design

The CCS will be the primary source of information that will feed into coverage measurement in the 2011 Census. It will be supported by other auxiliary information (such as administrative sources, information from the census field

operation), but the research into how that information is combined is at an early stage. However, since the CCS will be the primary source it should still be designed such that it could be used in isolation from the other sources. The CCS design must facilitate estimates of undercount for:

- Resident Households. It is important to ensure that we have an estimate of the number of households, thus we must measure how many the census missed.
- Resident Individuals. It is also important to be able to measure how these individuals are missed - either in missed households or in counted households to enable transparent adjustments based on this information.
- Key demographic variables e.g. age-sex, ethnicity. Precision of the age-sex estimates are the highest priority.
- Local Authorities.
- Cross-classifications of the key variables by area.

Target precision levels (for sampling errors only) are 95 per cent confidence intervals of 0.1 per cent around the national population estimate and 1 per cent for a population of half a million. Local Authority and age-sex level population estimates should aim for minimal variation of precision, therefore ideally being the same precision across all estimates. The intention is to try to deliver results that are better than the 2001 results, and in particular to ensure that there are no areas with a worse precision than the worst that was achieved in 2001 (i.e. there is no confidence interval for a Local Authority total population that is wider than 6.1 per cent). In fact, some improvement may be possible, and 5 per cent is perhaps an achievable upper bound.

Other principles which are guiding the CCS design are:

- The CCS should address the lessons learnt from 2001 (see Abbott and Brown, 2006).
- Simple methods should be developed where possible, to allow users to understand the methods behind the results. Gaining acceptance of the methodology from users is important. Users will not accept their census population estimates if they are not confident about the methodology used to derive them.

4. The 2011 CCS Design Strategy

Since the 2001 design was broadly successful, the overall strategy for the 2011 CCS will be similar, with a CCS of a comparable size and field methodology. It is clear that for robust coverage assessment the sample would still need to be area based as there will not be a household listing of very high quality that is independent of the census process (which will use a household listing that is checked in the field). Therefore, the basic CCS design is likely to follow the model adopted in the 2001 Census. It will be a stratified multi-stage sample of areal units that will be independently re-enumerated.

We need to decide the most appropriate sampling units and variables for stratification and clustering. This section outlines the sample design features where work is planned to explore improvements, including the choice of sampling unit, geographic stratification through forming Estimation Areas, further sub-Estimation Area stratification and the sample size.

4.1 Sampling units

The 2001 design used 1991 Enumeration Districts as its PSUs and postcodes as its secondary sampling units (SSUs). These choices were driven by the availability of data for stratification (1991 Census outputs which provided the data for EDs), their geographic coverage of the whole of England and Wales and survey practicalities (every household knows what postcode it belongs to so it is easy on the doorstep to determine whether a household is in scope).

For 2011, the data available is different. The 2001 Census used bespoke areas called Output Areas (OAs) as the lowest level of aggregation for most outputs. Each is a set of merged postcodes such that the population within them is relatively homogenous with respect to certain population characteristics. In addition, more non-census data is becoming available for Output Areas and their larger aggregations 'Super Output Areas' (SOAs). Initial studies have indicated that CCS designs that use OAs as their PSU are likely to provide the most efficient design for a fixed cost, despite their homogenous construction (it is normal to want clusters to be internally heterogeneous for an efficient clustered design). The internal homogeneity results in a high intra-cluster correlation of observational units within the OA, but we can recover the loss in efficiency by sampling fewer postcodes within the OA and selecting more OAs in the first place (although the total number of postcodes sampled is slightly fewer due to increased travelling costs).

In addition, as the estimation method we will use is based upon capture-recapture methodology, which assumes that the population subgroup to which you are applying it has homogenous response probabilities, then sampling areas that are internally homogenous is advantageous in terms of reducing the likelihood of heterogeneity bias in the estimation process.

4.2 Estimation Areas

One CCS objective is to provide the information necessary to measure undercount down to Local Authority level. In order to achieve this, it is sensible to build LAs into the survey design. If we stratified ignoring LAs, the estimation process would be more complex, harder to explain to users and therefore less acceptable to them. In 2001, the main geographic stratification came from forming Estimation Areas (EAs) – contiguous groupings of Local Authorities with a population of around 500,000 persons. For 2011, it would again appear sensible to form some kind of Local Authority grouping for the same reasons.

Some contiguous groupings used in 2001 were very heterogeneous because the geographical layout of England and Wales meant that in many cases urban areas had to be grouped with rural areas. This caused some difficulties at the estimation stage where adjustments were spread across the area, which needed a contingency approach to be used to ensure estimates were plausible. We are hoping to avoid the need for this, and perhaps to improve the efficiency of the estimates by grouping Local Authorities non-contiguously by some form of area type indicator, rather than restricting the groups by geographical constraints. Initial research shows almost no impact on the variances (rather counter to our expectations), and we are following this through to try to understand why this is so.

Contiguous LAs are attractive from a processing and survey management perspective, as 'blocks' of LAs can be built into the field management structure, and can perhaps be more easily prioritised in data capture (although technology may provide other solutions for this, at the risk of added complexity).

4.3 Stratification

From previous experiences, we know that undercount in a census varies by geography and demography. Therefore, to ensure our sample design is efficient, it is prudent to consider stratifying our sample to make use of this knowledge. Two types of stratification designed to reduce variability and ensure balance in the sample will be explored.

a) Socio-economic stratification

In the 2001 CCS design, the Hard to Count index was used to stratify the sample to increase its efficiency by controlling for variables that were, prior to 2001, expected to be correlated with the level of undercount in the 2001 Census. This index was derived using the 1991 Census data at 1991 Enumeration District level, since that was the only available source of rich low level data at that time. However, one of the weaker areas of the 2001 CCS as described in section 2 was its reliance on this 1991 Census information, as it did not reflect the variance it was designed to control. In some areas this meant the sample selection was not robust and resulted in estimates with either high variability and/or biases. This is an area of the CCS design that needs addressing, and ONS is continuing to examine patterns of response and to look for better and more up-to-date predictors for 2011 patterns of non-response to help form a stratification that is less prone to such problems. Also, if we are confident in our predictions then we could use a finer stratification than that used in 2001 to increase efficiency (although this increases the risk of balance problems occurring again). For instance ONS (2006b) describes a 5 level stratification that has been developed for stratifying the 2007 Census Test.

b) Demographic stratification

A simple way of reducing within stratum variability is to stratify the sample by a 'size' measure to ensure that the sample contains a good representation of the key population sub-groups for which we expect the undercount to be high. In 2001, key age-sex groups that were expected would have large undercounts were used in the design to form size strata. For 2011, the same approach could be adopted, although with greater knowledge of undercount patterns, the choice of variables could be improved. For instance, as well as ensuring balance across key age-sex groups, particular ethnic populations or tenure groups could also be included. In addition to using additional variables, improvements are being sought through access to updated population counts, to reduce reliance on using 10 year old census data in the design. ONS does produce small area inter-censal population estimates (although not by ethnicity or tenure), and these along with other alternative sources will be considered.

4.4 Sample size and allocation

It is anticipated that for 2011 the total sample size would be broadly similar to that in 2001. There are two main reasons for this. Firstly, there are always budgetary constraints and given the knowledge gained from 2001 it is expected that the same sample size will be required to provide estimates that are at least as good as those produced in 2001, one of the objectives of coverage assessment in 2011 described in section 3. If we are able to make significant gains in efficiency through some of the improvements discussed in this paper, it may be possible to reduce the sample size. This would be attractive if that size reduction could lead to an increase in CCS coverage and quality. However, this would have to be balanced against the risk of making the design less robust. Secondly, whilst the 2001 CCS was successful it would be risky to assume that the same success could be applied to a bigger survey. Therefore it would be reasonable to assume that the survey is not likely to be larger than the 2001 CCS, unless research shows that the expected precision in 2011 will be unacceptable. Therefore, the working assumption at present is that we will sample around 300,000 to 325,000 households.

For the 2001 CCS the sample was spread across all domains – i.e. geography, demography etc. The main reason for this was that it was the first time such a survey had been done on this scale, and there was little confidence in knowledge of

undercount patterns. Therefore, all Local Authorities had some CCS sample and the allocation of the sample across hard to count groups was fairly even. For 2011 there is much greater knowledge of the undercount pattern and its expected variability. Therefore the sample can be more targeted, as it would be in any sample design where we have prior knowledge about expected variability and other associated parameters. Thus a smaller sample size would be adequate in areas/domains where undercount variability was expected to be low and this would free up larger sample sizes for areas/domains where the undercount variability was expected to be higher. However, this does not necessarily mean a pure 'optimal' allocation - we will still aim for a robust design with some minimum sample size constraints. Demonstrating that smaller sample sizes will not unfairly disadvantage 'easy' areas will be critical to gaining user acceptance, and the constraints on the design will help in providing reassurance.

5. Evaluation of the design options

The previous section has highlighted that there are a number of options for improving the CCS design. These options need to be evaluated to explore the trade offs between gains in precision against increased cost and/or risk. In order to do this, two approaches will be used.

The first approach uses the 2001 Census results to derive estimates of sampling variances (including intra-cluster correlation coefficients when a two stage design is explored) assuming a simple expansion estimator to estimate the undercount. Then standard sampling formulae for the expected precision from the sample design can be applied (for example see Lohr (1999)). This can be applied to a variety of alternative stratification designs (e.g. different ways of grouping Local Authorities to form Estimation Areas) to enable precision comparisons for these alternatives. This indicates which options are most promising and should therefore be explored in more detail. The advantages of this method are that the data are available and it is not dependent on additional development work. The disadvantages are the number of simplifying assumptions made, making results difficult to interpret, and the levels of precision predicted are not realistic since they ignore the presence of the census as a powerful auxiliary.

The second approach uses a series of simulation studies to generate censuses and CCSs from individual level data. The structure of these studies is similar to those described by Brown et al (1999). Different sample designs (and estimation processes) can then be applied to examine the impact on the population estimates. The critical advantage here is that the patterns of coverage in the census and CCS can be varied to explore whether any design is robust to different scenarios. The disadvantage of these studies is that they are computationally complex and time consuming.

6. Summary

This paper has outlined the plans for the development of the 2011 Census Coverage Survey design. This work is critical, since the measurement of coverage relies heavily upon the sample design providing robust information from which the estimates of undercoverage are derived. The paper has highlighted the areas where improvements and opportunities are being sought, the most important being revisiting the stratification and the building in of newer intelligence and up-to-date data sources.

7. References

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

Brown, J.J., Buckner, L., Diamond, I.D., Chambers, R. and Teague, A. (1999). A methodological strategy for a one number census in the UK. Journal of the Royal Statistical Society, Series A, 162, 247-267.

Lohr, S.L. (1999) Sampling: Design and analysis, Duxbury Press, Pacific Grove, CA, USA

ONS(2003) Local Authority Population Studies: Full Report. Available at www.statistics.gov.uk/downloads/theme_population/LAStudy_FullReport.pdf

ONS (2006a) 2011 UK Census Coverage Assessment and Adjustment Strategy, Advisory Group paper (06)16. Available at www.statistics.gov.uk/census/pdfs/ag0616.pdf

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