

# **2004 Local Authority studies**

---

## **Adjustment Methodology**

---

## 2004 Local Authority studies: Adjustment Methodology

### Introduction

This paper details the methods used for adjusting the census results as part of the Local Authority Population Studies.

These studies were designed to identify areas where there was a risk that the census based population estimates could have been underestimated (or possibly overestimated), and the determination of a methodology for adjusting the estimates. The strategy and framework underpinning the studies can be found at [www.statistics.gov.uk/downloads/theme\\_population/LAStudy\\_FullReport.pdf#page=10](http://www.statistics.gov.uk/downloads/theme_population/LAStudy_FullReport.pdf#page=10)

The studies assessed whether there was sufficient evidence to justify an adjustment to the census estimates, based on the levels of risk and reliability. The findings of the studies can be found at [www.statistics.gov.uk/downloads/theme\\_population/LAStudy\\_FullReport.pdf#page=13](http://www.statistics.gov.uk/downloads/theme_population/LAStudy_FullReport.pdf#page=13)

Methods for adjusting the census results in light of these findings were developed within the One Number Census (ONC) framework by including a modified approach that allowed for specific local circumstances. These were designed to provide better estimates of population that used the evidence relevant to each Local Authority (LA).

The development of the adjustment methods focused on the assumptions that underpin the ONC process, such as the quality of the Census enumeration, balance of the Census Coverage Survey (CCS) sample and independence of the Census and CCS. Where there was evidence to suggest that the assumptions were not valid, more robust estimates of population were produced using the ONC framework, but including a modified approach to allow for these model failures. Further details of the ONC methodology can be found at [www.statistics.gov.uk/census2001/pdfs/oncguide.pdf](http://www.statistics.gov.uk/census2001/pdfs/oncguide.pdf)

### Background

The ONC methodology is based upon a number of assumptions that include the following:

- a) That the census was sufficiently free of localised poor enumeration or other 'non-random' issues (e.g. processing problems) for which the ONC method was not designed to compensate.
- b) Census response would generally be no worse than in 1991 when the response rate was 96 per cent. The levels of risk of underestimation would increase with the levels of undercount in the 2001 Census, as the ONC would have to estimate a larger undercount than that for which it was designed.
- c) The patterns of undercount would be similar to that in 1991, and if not, the ONC would be reasonably robust to the observed pattern.
- d) The Census Coverage Survey would find a high proportion of the households and persons missed by the Census. Furthermore, the level of dependence between Census and CCS would not be extreme and therefore together with (a), the impact of any dependence would be small.
- e) There would be high quality matching between Census and CCS, with a false negative rate of 0.1 per cent. (False negative matches are 'missed matches' - where two records which really do relate to the same household/person are incorrectly not matched).
- f) The ONC was designed to produce the best possible population estimates for Estimation Areas (EAs). Estimation for Local Authorities was possible but not directly, which led to the assumption that differences in Local Authority undercount were constant across age-sex groups within the hard to count (HtC) groups.

These assumptions, if violated, could have led to problems with the application of the ONC and a risk of bias. Therefore, within the Local Authority Studies, attention was focused on analysing the robustness of these assumptions and the likely impact of these on the estimates. In parallel with the analysis of the evidence, a series of methods were developed and tested that would address failures in these underlying

assumptions should there be evidence to suggest that these had occurred. These methods provided more robust estimates of population using the ONC framework, but included a modified approach to allow for the identified model failures.

The approaches developed, and the situations in which they might be applied, included the following:

**Method 1** - Estimating for localised poor response/enumeration failure - used where there was poor enumeration, high non-response or processing issues (including non-processed forms or a high number of late returned forms), for which the ONC had not compensated.

**Method 2** - Borrowing Strength - used where there was evidence that the CCS sample size was too small or CCS fieldwork was not of a high enough standard.

**Method 3** - Post-stratification - used where the original HtC index was not an efficient stratifier.

**Method 4** - Modified Outlier Strategy - used where the original outlier strategy was not appropriate.

These methods were designed to be complementary, so that where more than one issue was identified for a particular area, a combination of methods was able to be applied. Full details of these methods are provided below.

### **Method 1 - Localised poor response/ enumeration failure**

#### **Introduction**

One of the assumptions underpinning the ONC is that the undercount is randomly distributed within the strata used for sampling and estimation. That is, within a HtC and age-sex stratum the probability that a person is counted is not determined by their characteristics.

Violation of this assumption leads to model failure - and the resulting estimates of population will not adequately reflect the true undercount.

Such a violation could occur when, for instance, the quality of Census fieldstaff varies considerably. The CCS may not have measured this additional source of variability adequately at low levels, i.e. Enumeration Districts or wards. To protect against such a difficulty, the ONC QA process gathered all the available information from the Census fieldwork and management

processes so as to detect, and if necessary investigate, any quality concerns. However, the information was itself of a poor quality and was not robust enough to enable full analysis to be carried out.

However, the Manchester and Westminster matching studies highlighted that there were areas where enumerators had either failed to identify and deliver forms to all households, or areas where response was exceptionally poor. A study was therefore undertaken to examine whether there were other occurrences across the country, how to detect them, establish their interaction with the ONC process and consequently develop a method for estimating the population count within the ONC framework.

The following section outlines the methods developed for detecting and modifying the population estimates for localised poor response or enumeration failure.

### **Methodology**

#### **(i) Detecting areas of poor response/ enumeration failure**

In order to detect areas where there was poor response or enumeration failure, information about the numbers of households in a particular area was required. Council Tax (CT) data for 2001 at postcode level was utilised to identify both issues. However, there were known coverage and quality issues with Council Tax data for this purpose and therefore some care was taken to ensure the analysis was robust and consistent. The level of aggregation chosen for the analysis was wards - the reasoning behind this was as follows. Firstly, census enumerators worked within ward boundaries; secondly this level of aggregation should be sufficient to be able to detect poor response or enumeration failure; and thirdly the Council Tax data should be reasonably robust (in terms of location quality) at this level.

For the purpose of making these calculations possible at the ward level, the Council Tax was taken as a reasonable approximation of the true count of households. The implied assumption was that a Council Tax address would be equivalent to a single household. The returns from the Census fieldwork were then compared to these totals to calculate a response rate. Since we are only interested in occupied households, the Council Tax count was reduced by the observed count of Census vacants, second homes and visitor-only households to provide a count

of the number of residential households. This was then compared with the achieved census count of households to derive a response rate for each ward as given by (1), where:

- The subscript  $w$  denotes the ward of interest
- $returns_w^{CEN}$  denotes the number of returned census forms containing residents for ward  $w$
- $CT_w$  denotes the Council Tax count of dwellings for ward  $w$
- $vacant_w^{CEN}$  denotes the number of vacant households identified by census enumerators in ward  $w$
- $second_w^{CEN}$  denotes the number of second households identified by census enumerators in ward  $w$
- $visitor_w^{CEN}$  denotes the number of returned census forms with no residents but containing visitors in ward  $w$

The next stage was to determine whether the ONC process was able to make a robust adjustment for undercount across all wards. The ONC may not have been able to correctly compensate for the undercount for the Local Authority as a whole if either there was a large variability in the response rates or there were clearly outliers in the distribution.

In order to determine whether this was indeed the case, the overall Local Authority household response rate measured by the ONC was compared to the ward level distribution of response rates arising from (1).

If the council tax response rate was 10 per cent lower than the corresponding ONC response rate within a ward, then this ward was considered at risk with respect to the ONC adjustment. e.g. if the ONC household response rate was 71.5 per cent any wards with a Council Tax implied response rate of below 61.5 per cent would be extracted and further investigated.

Wards identified using the 10 per cent criterion were then examined in more detail, to explore whether there was evidence from field information to confirm that there were either enumeration problems or extremely

poor response. This evidence was gathered by examination of the Enumerators Record Books, and cross referencing the information in those books with lower level information from the Census database and the Council Tax records. This was used to provide information on the quality of the Council Tax data, especially in relation to information about derelict or demolished properties (which may be on the Tax file but not in the Census) or Communal Establishments (where a single Census record may relate to many Tax records such as a caravan park).

Further investigations researched the location and performance of the CCS sample. This was also important, because this would provide evidence to show how these potential ward level problems affected the One Number Census estimates. The CCS sample sizes within the wards examined were important - if there was no sample or the sample size was very small, then this increased the likelihood that a robust adjustment was not possible in the ONC. If there was sample, its performance (i.e. whether there was evidence to suggest that the CCS found an acceptable proportion of persons and households missed by the census) was compared with that implied by the Council Tax - if the CCS was not finding the missed households implied by the CT then again this increased the likelihood that a robust adjustment was not possible in the ONC.

These investigations established confidence in either the Council Tax figure (suggesting an adjustment would be needed) or the Census household figure (suggesting an adjustment might not be needed). If the investigation explained some of the difference, thus reducing it to an acceptable level, it could be judged that there were not any issues of localised undercount in that ward and it could remain in the ONC estimation process.

#### (ii) Adjusting the ONC to account for areas of poor response/enumeration failure

The wards identified using the above criteria were considered as being beyond the scope of the ONC to make a robust adjustment. There was no CCS sample in most of these wards. Therefore, the solution is to post-stratify into

$$(1) \quad \text{responserate}_w = \frac{returns_w^{CEN}}{(CT_w - vacant_w^{CEN} - second_w^{CEN} - visitor_w^{CEN})}$$

two new strata - a stratum where the ONC still operates, and a stratum where the population of the identified wards are estimated using a third source of information.

To achieve this, the identified wards were excluded from the ONC estimation process, and a new ONC estimate for the LAs less those wards was produced using the regular unchanged ONC processes. This estimate was (generally) lower than the original ONC estimates, as the auxiliary variable (the census counts) used for prediction is lower due to the removal of the wards. Note that if there was CCS sample within the identified wards, this sample information was also removed. This could affect the estimated coverage patterns in either direction.

The population estimate for the ward(s) in question was then calculated using the difference between the number of Census households and the Council Tax less vacant and second homes figure for the ward as shown in (2). This provided an estimated number of households missed by the Census.

To convert this into persons, this difference was multiplied by the household size from the existing census households (excluding vacant households and second homes) in the relevant ward as shown in (3). This gave an estimate of the total number of persons missed in the Census by ward.

Age-sex estimates were obtained using the demographic profile for the ward from the counted Census person information applied to the estimate of 'missed' persons. This has the advantage that it uses the demographic profile across the relevant ward. Hence, any localised characteristics are taken into account, but it is assumed that the characteristics of 'missed' persons in each identified ward are the same as those that were counted.

The ward estimates were then added to the new ONC estimate to obtain a new whole LA estimate. The difference between this and the original ONC estimate was the adjustment.

## Method 2 - Borrowing strength

### Introduction

For the ONC estimates of population to be robust, there must have been sufficient CCS sample on which to base adjustments and the survey fieldwork must have been successful. If either of these two prerequisites were not achieved, the population is at risk of being under-estimated. During the original ONC process, the small sample size issues were mainly solved by collapsing strata. It should be noted, however, that since estimates were computed within the design strata the two pre-requisites did not generally present an issue). Furthermore, the qualitative information that was available at the time indicated that the CCS was on the whole successful.

However, the Local Authority Studies reviewed all evidence for a particular area, and in some areas demonstrated that the sample sizes were insufficient. This is especially the case if the HtC index based on 1991 data did not reflect the true characteristics of an area in 2001 or if the performance of the CCS was less successful than previously thought. In these cases, the ONC estimates were essentially rejected. In most cases this was for a particular stratum, such as the HtC 3 stratum. However, there were some cases where no other data sets were considered reliable and consistent enough to be substituted for the rejected ONC estimates across a whole stratum. The contingency was to use information about similar LADs that had already passed quality assurance to make adjustments to the rejected ONC estimates. This process is called 'borrowing strength' and was based on the original contingency strategy from the ONC process (see section 5 of [www.statistics.gov.uk/census2001/pdfs/oncinfopaper.pdf](http://www.statistics.gov.uk/census2001/pdfs/oncinfopaper.pdf) for details of the methodology and the selection of similar areas).

### Methodology

If the CCS did not work as well as expected, not enough additional people (missed in the census) would be added through the ONC process, and therefore the implied response rate would be too high. The borrowing strength areas, where the CCS functioned as expected, therefore had lower

$$(2) \quad \text{householdsmissed}_w = CT_w - \text{vacant}_w^{CEN} - \text{second}_w^{CEN} - \text{visitor}_w^{CEN} - \text{counted}_w^{CEN}$$

$$(3) \quad \text{persons missed}_w = \frac{\text{householdsmissed}_w \times \text{personscounted}_w^{CEN}}{\text{householdscounted}_w^{CEN}}$$

response rates. Using the ONC adjustments for the borrowing strength areas to the particular area increased the population estimate. The borrowing strength areas have already been defined through the ONC consultation process. The method takes the unweighted mean response rate across the Hard to Count and age-sex groups of the borrowing strength areas, and applies this response rate to the stratum where the original ONC estimate was rejected. Equation (4) shows this calculation, where  $a$ ,  $h$  and  $l$  are the age-sex group, the hard to count group and the LAD for which we are borrowing strength and  $n$  is the number of LAs from which we are borrowing strength. The factors are simply the ratio of the ONC estimate to the Census count defined by (5).

$$(4) \text{ estimate}_{ahl}^{BS} = \text{census}_{ahl} \left( \frac{1}{n} \right) \sum_{i=1}^n \text{factor}_{ahi}$$

$$(5) \text{ factor}_{ahi} = \frac{\text{estimate}_{ahi}^{ONC}}{\text{census}_{ahi}}$$

If an area did not have any adjustment in a particular HtC stratum (e.g. as it is collapsed) then it was not used for that stratum. Similarly, if one of the borrowing strength areas was one for which an adjustment was made, it was not used within this strategy (e.g. Westminster was not used as a borrowing strength area for Wandsworth because an adjustment was made in Westminster).

### Example

The HtC 3 stratum for a particular LA does not have enough sample to make a robust ONC estimate, therefore the borrowing strength strategy is invoked. There were 37 age-sex groups used by the ONC process, so a total of 37 separate estimates had to be re-calculated. For each of the 37 age-sex groups within the HtC 3 stratum, the adjustment made to the Census figure was obtained for each of the five borrowing strength areas within the same stratum. For instance, within one of the five borrowing strength LAs if the Census counted 4000 people but the ONC estimate was 4400, the adjustment factor would be  $4400/4000=1.1$ . The mean adjustment over the five LADs was then calculated for each age-sex group in the HtC 3 stratum. This adjustment was then applied to the census count for the LA where the estimates were rejected. For example, if

the Census count for an age-sex group within the HtC 3 stratum of the rejected LA was 5000 and the mean adjustment factor was 1.08, the estimated count would be  $5000 \times 1.08 = 5400$ . The new estimates for HtC 3 were then aggregated with the original HtC 1 and 2 estimates that had not been rejected to produce estimates for each of the 37 age-sex groups. These were rounded to the nearest whole number where necessary. This provided a new set of population estimates.

## Method 3 - Post-stratification

### Introduction

The ONC estimation process used pre-defined stratification. These were the Hard to Count index and 37 five year age-sex groups. The Hard to Count index was a national small area classification based on 1991 Census data, and was used both to draw the sample and to stratify for estimation. Further information on the research underpinning the Hard to Count index can be found at [www.statistics.gov.uk/census2001/pdfs/sc0015.pdf](http://www.statistics.gov.uk/census2001/pdfs/sc0015.pdf). Whilst this was the best stratification that could be used at the time, there were areas where the Hard to Count index may have been a poor stratifier. If an area had undergone substantial redevelopment or change since the 1991 Census, the sample selection may not have adequately reflected the true population distribution. As part of the Local Authority studies, analyses were undertaken to examine the change between the 1991 Hard to Count distribution and that implied by the 2001 Census data.

Post-stratification was used in the situation where there was evidence to suggest there were problems with either the balance of the CCS sample across the HtC groups (where, for instance the overall sample was adequate but it was unbalanced within a HtC stratum) or where the 1991 based stratification was found to be an inefficient stratifier for estimating undercount (where, for instance, the hardest to count areas had changed significantly since 1991).

Post-stratification is a standard method for estimating population parameters within a newly defined stratification, and it uses the newly collected data. The 2001 Census data were used within the ONC framework to define a new Hard to count index. This new index had the advantage of being more up to date and the potential of being a better stratifier for estimating undercount in areas that had changed significantly. However, post-stratification also

implied that the sample would be split into different strata - and that all areas would be re-allocated for prediction. Therefore the achieved sample sizes were sometimes small in the new HtC post-strata, and for this reason there was no guarantee that the post-stratification was better. The differences between a post-stratified estimate and the original Hard to Count index were not always large.

This strategy was used a few times within the original ONC process, where the 2001 Census data was used to construct a new hard to count index using the same variables. However, this was only possible at a local level, and it was found not to make substantial differences to the estimates of population.

However, for the Local Authority Studies it was possible to use a nationally based post-stratification. The new HtC index used 2001 Census data for the same variables included in the original index, constructed at a national level. Therefore this new index was nationally consistent and more robust than the localised post-stratification used in the original ONC process.

### Methodology

A hard to count score was calculated at the postcode level from the sum of the proportions of 2001 Census variables for:

- Multioccupancy (Approximated using shared dwellings in converted accommodation)
- Private Rented Households
- Unemployed Persons
- Non-responding dummy forms
- Non-native English speaking defined by Country of Birth

A three level hard to count index using a national 40 per cent, 40 per cent, 20 per cent distribution was constructed from the ordered scores. Thus 40 per cent of postcodes with the lowest scores nationally were assigned to hard to count index level 1. This was identical in construction to the original Hard to Count index, although it used up to date data.

The newly constructed index was used within the original ONC estimation process, replacing the previous index. Both the CCS and the Census data used this new stratification.

## Method 4 - Modified outlier strategy

### Introduction

The One Number Census estimation strategy used a combined dual system and ratio estimator at Estimation Area level to derive population estimates. Ratio estimation can be sensitive to extreme observations, and several modifications were made to the standard ratio estimator to make it robust. More information on these modifications can be found at [www.statistics.gov.uk/census2001/pdfs/sc0003A.pdf](http://www.statistics.gov.uk/census2001/pdfs/sc0003A.pdf) and [www.statistics.gov.uk/census2001/pdfs/sc0016.pdf](http://www.statistics.gov.uk/census2001/pdfs/sc0016.pdf).

One issue were cases where the CCS found a large number of 'missed' persons within the same postcode and age-sex group, that were also considered as outliers. Outliers, as defined within the ONC strategy, were identified by using pre-defined 'ratio' limits for individual observations. The ratio was defined by the ratio of the dual system estimate (DSE) to the census count for each age-sex observation in each postcode. These limits differed for each Hard to Count stratum. The limits were a ratio of 3 for the Hard to Count 1, a ratio of 4 for the Hard to Count 2 strata and a ratio of 5 for the Hard to Count 3 strata.

For example, if within a single HtC 3 postcode the CCS found 6 males aged 20 – 24 that the Census had missed and the Census and CCS both found 1 male aged 20 – 24 (and there were no males aged 20 – 24 in the Census that the CCS failed to find), then the dual system estimate for that postcode would be  $(1*7)/1=7$ . The ratio of the DSE to the census count is 7, which is greater than the cut off ratio of 5 for the HtC 3 stratum, and therefore this observation would be classified as an outlier.

If an observation was classified as an outlier through this method, it was not used in the calculation of any model parameters. Instead, it contributed to the population estimate by using the DSE as its predicted value. That is, it was not used to represent any other areas but did contribute its DSE to the population. This methodology was based on the data from the simulation studies, assuming an overall 95 per cent Census coverage. The strategy was therefore designed to identify rare situations for which it was implausible to infer that the same situation had occurred in many other locations within the same Estimation Area. However, in situations where census response was much lower, this strategy introduced a risk of underestimation, because it removed observations for which it was

plausible that similar occurrences were present elsewhere.

### **Methodology**

For Estimation Areas where there was evidence to suggest that the outlier strategy was not appropriate, a modification was made to the ONC estimate process, in which the outliers identified in the original ONC methodology were not excluded from the estimation of ratio parameters. Instead, they had their influence reduced to the predefined cut-off values used to identify outliers - a method known as winsorisation, a practice commonly employed in survey estimation. The 'outliers' were therefore included in the usual estimation process, thereby increasing the estimates of undercount. Using the example above, the outlier would not be excluded, but the observed DSE would be reduced from 7 to 5, and the observation included in the ratio estimation processes.