national STATISTICS

Census 2001 data quality:

Variability in tabular counts

2001 Census data quality: Variability in tabular counts

1 Introduction

No large scale data collection exercise will ever be 100 per cent accurate and we would expect some variability in the final published results. This means that there might be small differences between the 'true' counts in a population and the estimated counts that are published. The 2001 Census is no exception and there are many sources for variability to occur for person, household and communal establishment tables.

This paper sets out to describe some of these sources and where possible, presents a guide to the level of variability that may occur within published tables at different geographical levels.

2 Sources of variability

The main sources of error can be categorised as coverage, respondent and processing errors:

Coverage error occurs as a result of missing, overcounting or incorrectly including dwellings or individuals in the census.

Respondent error is where incorrect information may sometimes be given on the census form or questions are unanswered.

Processing error can occur during several stages. Although the following processes are in place to improve the data, an element of variability is introduced. The main processes that may add variability to the data are:

i) **One Number Census** where census data are adjusted for underenumeration by imputing individuals and households back into the data

ii) **Data capture and coding** where information is scanned, keyed and coded from the returned census form

iii) **Edit and imputation** where inconsistent responses are adjusted and non-response is removed

iv) **Disclosure control** where adjustments have been made to the data in order to protect the confidentiality of information.

2.1 Coverage error

A census differs to a sample survey as it aims to enumerate every member of a defined population. Despite best efforts in identifying dwellings and following up the non-return of Census forms, there will always be problems of under-coverage and over-coverage. An adjustment of final census counts is made to account for under-coverage and there were procedures in place to address over-coverage by deleting repeated records. Over-coverage may also occur when an individual appears on more than one census form or an individual is incorrectly included as a usual resident when they are a temporary visitor.

2.2 Respondent error

Sometimes incorrect information may be given on census forms or questions left unanswered. Inaccurate information may be identified as inconsistent (such as a three year old with degree qualifications) or it may be feasible (such as a married 23 year old instead of a single 23 year old). Where there was inconsistent or missing information, a process amended the data through the statistical technique of imputation.

In addition to changes made by the process there remains variability in the results caused by respondent error that is highly problematic to measure. There was a degree of contact with some respondents after Census day to confirm key variables such as age, sex and marital status, but it was not feasible to check all the information provided on all Census forms. The Census Quality Survey (referred to later) was a useful mechanism to identify common causes of respondent error.

2.3 Processing error

The One Number Census (ONC) aimed to ensure full coverage of the total population. It incorporated the use of a census coverage survey to make final estimates of the population. Where under-enumeration had occurred, the system imputed full person and household records using a donor system. More information of the ONC methodology can be found at **www.statistics.gov.uk/census2001/onc.asp**.

The transfer of information from a paper form to an electronic format is a large and technically intensive process. Information was either electronically captured or manually keyed. Numerical codes were assigned to some variables automatically and some variables required manual coding by trained staff. The process of data capture and coding was subject to error and its accuracy was regularly monitored. The quality checking process examined questions independently but did not check for consistency of responses across the census forms, as inconsistencies would be identified in a later process. The results of the quality checks showed that generally the quality of the data capture was very high. For further information, please see **www.statistics.gov.uk/census2001/proj_proc.asp**

The Edit and Donor Imputation System (EDIS) was designed to correct missing and inconsistent information on census forms. The process located a donor record according to a set of key variables and imputed (copied) information from the donor to the record that had the missing or inconsistent response. The result is a valid, but not necessarily correct response. The system adopted a principle of minimal change where the smallest amount of information on the census form was altered. The project evaluation report can be seen at **www.statistics.gov.uk/census2001/proj_eai.asp**

Disclosure control is a process that purposely adds uncertainty to Census data. ONS cannot release any data that may allow identification of individual person or household records. Record swapping and small cell adjustment have been used to protect information and these add uncertainty to tabular counts, particularly small counts. More information on disclosure control methods is available at **www.statistics.gov.uk/census2001/discloseprotect.asp**

3 Quantifying the variability

Although we know that the above sources of variability exist, they are not easy to quantify. Respondent error and edit and imputation accuracy are very hard to measure as it can be difficult to find out what the "true" values are. However, we do have information on response rates, accuracy of data capture and coding processes and disclosure control adjustments. Additionally, a study was carried out before the 2001 Census to investigate the effectiveness of the planned edit and imputation system by putting 'holes' into some data for several variables. Combining these sources of information will give a reasonable indication of the variability present within the tabular counts.

3.1 Coverage

The final estimates are open to sampling error because they are based upon a sample survey. The One Number Census (ONC) process was controlled at the local authority level and higher variability is expected at the ward and output area levels.

Information about the quality and variability of the ONC process has already been published. A 95 per cent confidence interval for the population estimated by the process for each local authority can be found at **www.statistics.gov.uk/census2001/downloads/95conf.xls.** Confidence intervals for wards would on average be larger than those at local authority level due to the smaller population size of wards. Similarly we would expect output area level confidence intervals to be larger than ward confidence intervals.

The ONC imputation rates by local authority and key variables are available at **www.statistics.gov.uk/census2001/imputation_rates_by_variable.asp**. The imputation rates have been used as a measure to assign a quality indicator to each ward and these can be found at **www.statistics.gov.uk/census2001/quality_indicators.asp**.

The variability added to the data from the ONC process is the largest measured source of variability present within the data. Data from a sample of 40 wards showed that the ONC process contributed about three-quarters of the total variability to the counts for most variables.

3.2 Respondent error

Respondent error that caused inconsistent responses in census data can be quantified to some extent by examining the records that have been amended by the edit and imputation process. However, the level of respondent error caused by incorrect but feasible responses is very difficult to quantify. The Census Quality Survey carried out in 1999 compared responses given on the census form with responses given in a follow up interview. The survey compared the responses and identified where the answer given on the census form was different from the answer given at the interview. The main reasons for differences related to the interpretation of

questions and definitions. Key variables such as age, sex, marital status and ethnic group had about 95 per cent or higher levels of agreement between the two answers given. More information resulting from this survey will be available shortly.

3.3 Data capture and coding

There were regular quality checks throughout the data capture process and electronic information was compared with images of census forms. Checks for consistency of coding were carried out where the forms were processed and accuracy assessments were made after the data had been delivered to ONS. These assessments involved checking the coding for a sample of records with independently coded data. The accuracy of the data capture varied depending on how data for specific variables were captured. The highest levels of accuracy were achieved by the Optical Mark Recognition software that captured tick box responses (such as sex). The lowest level of accuracy was for the industry variable that required automatic and manual coding. The accuracy of data capture and coding by type of data capture can be seen at **www.statistics.gov.uk/census2001/processingevrep.asp#key**.

3.4 Edit and imputation

Inconsistent responses were identified during edit checks, and values were then imputed that would maintain consistency with the remaining responses on the census form. Similarly, responses were imputed where a question was unanswered.

The edit and imputation process has been very effective in imputing realistic distributions of variables where inconsistent or missing responses have occurred. Although the process improves the quality of the data, it contributes some variability to the final counts. Therefore the variability added by the process will be greater where there are higher non-response rates. Response rates for questions by local authority can be seen at **www.statistics.gov.uk/census2001/downloads/ItemnonrespLAD.xls**

Comparisons of non-imputed and imputed distributions by each variable give a good guide for identifying any obvious errors that may have occurred by the process. We expect the distributions to differ to some extent because of the likelihood of non-response bias for some variables.

Although it was not generally possible to compare imputed and true values, people's names were used as a means of checking the accuracy of imputation for sex. In total, imputations were made for sex for 0.4 per cent of the population. The investigation (based on a sample of areas) showed that the non-imputed and imputed distributions were very similar and that there was 75 per cent accuracy for the imputation of sex. The remaining 25 per cent of values were imputed to the opposite sex of that suggested by the name field, but these records had the correct proportions of males and females imputed.

The results of this investigation can be seen more fully in the evaluation report for the project at **www.statistics.gov.uk/census2001/editimputevrep.asp**

3.5 Disclosure control

The variability added by disclosure control methods can be measured by comparing data before and after the measures are applied. The uncertainty from disclosure control is added by two methods: record swapping and small cell adjustment.

In the majority of cases, distributions of variables are not largely affected by record swapping. There is random variability attributed from record swapping that does not follow any particular pattern and sometimes there can also be a systematic shift in the distributions. Where records are swapped between wards that have different characteristics from those of the local authority district they nest in, there can be a shift in the ward distributions towards the local authority distributions. This tends to make wards within a local authority area more homogeneous (similar in characteristics). The proportion of records swapped is confidential in order to maintain the protection offered by the method.

Small cell adjustment adds some uncertainty to small counts in tables so that information about individual persons or households cannot be identified. The method is designed to be unbiased and the tables are designed to be internally additive (although some discrepancies may occur when comparing values across several tables or aggregating data for several areas). The variability introduced by this technique will be largest within data for small areas, that are likely to contain small counts. The effect of small

cell adjustment is difficult to quantify without disclosing details of the method itself, which must remain confidential. However, it is possible to examine differences of aggregated area counts (that are likely to have been adjusted) with total counts of the area they nest in (such as output area data aggregated and compared with ward level data). The unbiased methodology suggests that the differences would be zero and any non-zero differences are due to the random variability of the process.

4 Measuring the variability from processes

For this exercise, data from a sample of 40 wards were selected and the variability contributed from each source was examined. Four key variables were examined; age, sex, marital status and ethnic group. The variability of the different sources was measured as follows:

- Data capture and coding variability was measured using the accuracy information gained from the quality checks on the different types of data capture and coding.
- ONC variability was measured using an estimate of the ward level confidence intervals based upon the local authority level confidence intervals.
- Edit and imputation variability was estimated by comparing imputed and non-imputed distributions of variables and checking images of census forms.
- Disclosure control variability was obtained by comparing tables before and after the adjustments were applied.

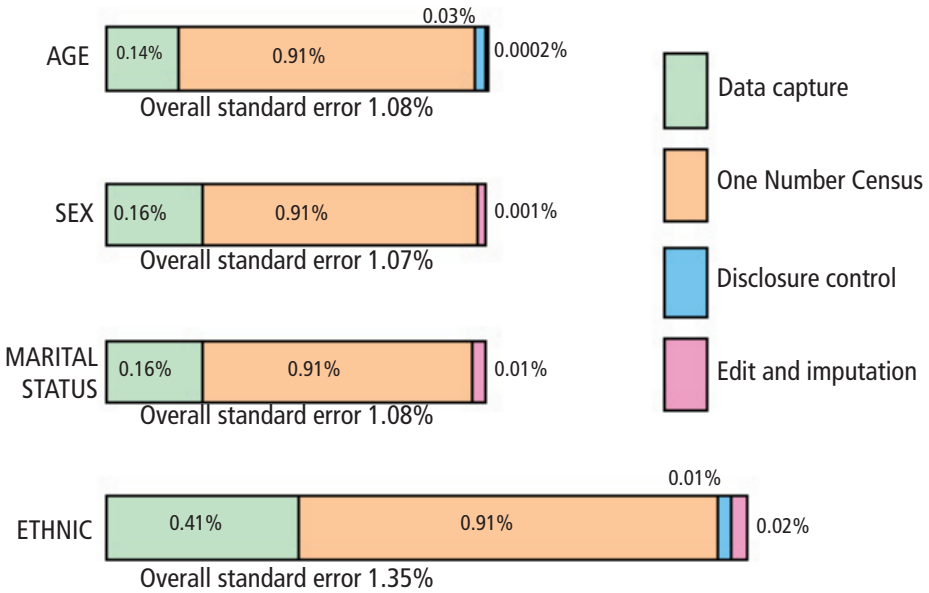
The figure below presents the results of the exercise. It uses the standard error statistic, which provides a measure of variability. The larger the standard error, the more variable the data.

Confidence intervals can be calculated by multiplying the standard error by 1.96 (if 95 per cent confidence intervals are required). Therefore a 95 per cent confidence interval for age is 2.2 per cent. This means that due to the variability within the data, on 95 per cent of occasions we would expect the true count to be within 2.2 per cent of the published count.

The graphic shows that the majority of the variability is from the One Number Census process and the variability from the disclosure adjustments is very small when compared with the other sources.

The work presented here has some limitations as it is based on a sample of data at the ward level of geography. Forty wards were chosen for convenience and a larger sample may have been beneficial to acquire a more accurate assessment.

Figure 1
Variability of tabular counts by processes



Note: because of the cell sizes, no small cell adjustment was applied to sex or marital status

We would expect larger proportions of the variability from disclosure adjustments and the One Number Census at the output area level than presented here. Generally, we would not expect the variability from data capture or edit and imputation to vary much by geographical level. However, the variability itself will change depending on the characteristics of different areas.

Further limitations are the methods of assessing accuracy for the imputation processes. The accuracy of data capture and coding and the variability of disclosure adjustments are easier to measure as they can be directly compared with the information available from images of census forms and unadjusted data. An assessment of imputation processes is largely based on judgement of how feasible the imputed distributions are compared to the non-imputed distributions. We would not expect the two distributions to be the same as there is a strong likelihood of non-response bias. This is where individuals with certain characteristics are more likely to leave a question missing or not return a census form. Therefore further investigation for non-response bias was carried out where the distributions differed.

5 Conclusions

As with any survey, there will always be variability from many different sources that will add some uncertainty to the published figures. This paper has looked at sources of variability in terms of respondents, coverage and processing. The variability of the data caused by processing has been investigated further in terms of data capture and coding, One Number Census, edit and imputation and disclosure control.

These processes have been very effective in capturing the data, improving its quality of use and protecting the confidentiality of information. The added variability is a small consequence of the improvements that have been made and the protection of individual information.

An analysis of data from 40 wards showed that the largest source of variability was from the One Number Census process that imputes whole persons and households where they have been missed from the Census. The variability is greater in areas of smaller population sizes and in areas of high non-response.

The next largest source of variability was from the data capture and coding process, where there were different levels of variability by question. Some questions required manual coding and the possibility for inaccurate coding is greater for these questions. The majority of the information was obtained from Optical Character Recognition and Optical Mark Recognition, which have very high levels of accuracy.

The process used to correct missing and inconsistent responses (the Edit and Donor Imputation System) attributed different levels of variability for different questions. The system was highly effective in imputing realistic distributions of categories of variables and variability in the data as a result of this process tends to be small compared to other sources.

Disclosure control is essential to protect individual information and it is the only process that is designed to add uncertainty to the data. The greatest levels of variability are expected within tables that consist of many small counts. These may be tables of small geographical areas, or tables of skewed distributions, such as ethnic group.