



Census 2001 Review and Evaluation

November 2003

Data Validation: Executive Summary

Content	Page
Project Objectives.....	2
Background.....	2
Methodology.....	2
Assessment and Lessons Learned.....	4
Lessons Learnt.....	5
Conclusions.....	6

ONS is carrying out a review and evaluation of the 2001 Census in England and Wales which will culminate in a Data Quality report and a General Report being published.

Plans for individual reports on specific aspects of the Census operation and a timetable for release have been published.

Each report is written in isolation and is subject to amendments as processing progresses and further information comes to light.

Reports will be released on the ONS website in the form of a high level Executive Summary and a more detailed Evaluation Report.

Census Customer Services
ONS
Titchfield
Fareham
Hants PO15 5RR

Telephone: ++44 (0) 1329 813800
Fax: ++44 (0) 1329 813587
Minicom: ++44 (0) 1329 813669
E-mail: census.customerservices@ons.gov.uk
Website: www.statistics.gov.uk/census2001

Census 2001 Review and Evaluation

Project Objective

To ensure census data could be used with confidence, by minimising the level of systematic error in the data.

Background

The data validation process was set up to carry out checks and where necessary to make corrections designed to improve the quality of census data. The concept of quality can be described as 'fitness for purpose' in terms of user needs. A strategy for improving quality is always a balance between the improvement gained and the time and resource required.

Initially, the project was set up to reduce the risk of users finding incorrect data following the types of coding error which were found in the 1991 census at the output stage. There was a need to establish systematic checks and validate data at various stages of processing of the 2001 Census.

Research was undertaken into how other countries approached Data Validation in order to identify best practice. It was found that they compared census data with secondary data derived from valid surveys or previous censuses. A Data Validation Strategy was then derived with the objective of ensuring that customers could use the 2001 Census Data with confidence, in particular by identifying any systematic errors at the earliest opportunity so as to avoid them appearing at the final output stage where corrections would be more costly and time consuming.

The key features of the Data Validation strategy were to:

- use secondary data from other surveys as a basis for validating the 2001 Census data;
- set up an independent Data Validation Team to validate the Census data at Local Authority (LA) level. The team's responsibilities included specifying the tolerances for which the checks would be made, validating data, referring unexpected issues/errors to an Issues Management group for guidance and checking the solutions;

- set up an Issues Management group to take responsibility for resolving any problems in the Census data;
- make use of Automatic Validation Checks (AVC) allowing checking to be more systematic and faster with a 'Pass', 'Fail' and 'Refer' to advise. This approach would give the Data Validation team greater time to investigate discrepancies and possible errors;
- cooperate with Northern Ireland Statistics and Research Agency (NISRA) and General Register Office for Scotland (GROS) in validating their data by sharing developed software tools and knowledge. NISRA and GROS were responsible for validating their own data. Different approaches were applied by the three offices in some areas; the balance with time and resource was different because of larger volumes of data in England and Wales;
- set up a Data Quality Monitoring System (DQMS) which would be used by the Data Validation team. The DQMS would be used to tabulate large volumes of data and then compare distributions with expected distributions. It would be able to dynamically interrogate large volumes of data, to allow queries down to record level to be made. It would then produce reports to identify whether the Census data had 'Passed', 'Failed' or had been 'Referred'; and
- undertake Validation at three critical stages of the projects lifecycle: at Load, after Edit and Imputation and after the data is ready for output.

Methodology

The initial validation methodology was based on the strategy of comparing census data with secondary data gathered from previous census and other surveys. Three processes were identified:

Set up

The initial stage prepared secondary data for comparison with census data, and defined criteria against which the Automatic Validation Checks (AVCs) would be carried out. 'Pass' was flagged when the Census data were regarded as acceptable, 'Fail' when there was

Census 2001 Review and Evaluation

a significant difference that required investigation and 'Referral' when there was no secondary data for comparison. A DQMS database was set up from which reports were generated outlining the relative number of passes, fails and referrals at the LA level.

Run

Census data was compared at three different points in the processing cycle: at Load (after the data had been captured and coded), after Edit and Imputation (where missing values had been estimated and the data for respondents was complete), and after the data was ready to be output (after the One Number Census process had added data for non-respondents). Each run culminated in a DQMS Report by LA.

Analysis

Once the DQMS report was complete for the LA, investigations were carried out on items that 'Failed' or were 'Referred'. The team could carry out SQL queries on the data and refer to images of census forms. Where a serious error was identified it was referred to the Issues Management group for a decision on the way forward.

Implementation

In carrying out this initial strategy several shortcomings were identified including:

- The amount of time taken to validate each Estimation Area (EA) was found to be excessive.
- Many of the AVCs did not have suitable secondary data on which to make a comparison.
- Sampling errors in the secondary data caused many apparent failures to be identified when the AVCs were run.
- The relative level of the failures of the AVCs was unknown with minor anomalies being investigated at the expense of more significant errors.

Revision of Strategy

A revised strategy was then implemented based upon looking at the 'big picture' in which smaller anomalies would not automatically be investigated as detailed checking of those minor anomalies was unlikely to add significantly to the quality of the data. This change in emphasis allowed priority to be given to potential major errors and the better targeting of resources.

In order to undertake this strategy, validation was carried out by EA, rather than at LA level, with comparisons being made variable by variable. The data for each EA was compared with the average and the range of data for previously analysed EAs. Trends could then be identified and a decision made on whether to undertake further investigation.

Implementation of new Strategy

The new strategy enabled each EA to be validated more rapidly. However, use was still made of several elements of the original strategy:

- Validation was carried out at the same three stages.
- The software tools and expertise that had been built up continued to be applied for investigations of errors.
- When errors were found use was made of the Data Quality Review Panel (DQRP) to discuss effective solutions.
- The EAs which had been validated using the original methodology provided initial comparative data that was used when loading the first EAs into the new system.

Census 2001 Review and Evaluation

Assessment and Lessons Learnt

In order to assess the success of the project, we provide examples of several key quality issues that were identified and resolved by the Data Validation strategy:

Black Lines

Unusual ticking patterns were noticed in some batches of forms. Image checks revealed black lines running through unticked boxes, mainly on even numbered pages of census forms. The problem was found to have arisen during the processing of forms; lines of dust sometimes settled on the scanners causing a black line to appear; where the line passed through a tick box, false information was recorded as if the box had been ticked. The variables mainly affected by black lines were religion, ethnic group, qualifications and relationships.

The contractor, Lockheed Martin, revised their cleaning procedures, which significantly reduced the occurrences of black lines. They also revised their procedures for checking images. By the time these additional procedures were introduced, some three quarters of census forms had been scanned. To fix the problem for already scanned forms, research was carried out and automatic corrections were applied to the religion and qualifications questions. Fixes were applied to batches of forms where unusual ticking patterns were found by Enumeration District (ED). In total, 36,652 amendments were made for professional qualifications, 61,425 amendments for educational qualifications and 99,949 for the religion question. Full details of the black lines problem are provided in the Quality of Data Capture and Coding report [www.statistics.gov.uk/census2001/pdfs/data_capture_and_coding_evr.pdf]

Armed Forces

Initial comparisons suggested a considerable apparent undercount of both UK and foreign Armed Forces personnel. This was confirmed when census figures were compared with data provided by the Defence Analytical Services Agency (DASA). The comparisons showed that the Census count fell within 5% of the UK Armed Forces figures provided by DASA for only 13 out of 376 LAs in England and Wales.

Research identified several contributing factors:

- Foreign Armed Forces were incorrectly coded to UK Armed Forces.
- Members of the Armed Forces were asked to state their rank instead of their occupation. Those who did not give their rank were coded to their equivalent occupation and were not included in the Armed Forces counts.
- Some members of the Armed Forces had correctly been coded to their rank but the Industry coding was incorrect, again resulting in them being excluded from the Armed Forces counts.
- It is also likely that some Armed Forces personnel had both Occupation and Industry incorrectly coded.

To establish high quality armed forces data, a separate database was set up containing people believed to be members of the armed forces. A list of potential armed forces personnel was extracted, based on a set of specified criteria. The text data for 'Job title', 'Job description', 'Business of employer' and 'Organisation' for these people were then matched against a list of phrases created from text data for people who had already been coded as members of the Armed Forces. All people who had a positive match against either Occupation or Industry text were added into the database for Armed Forces personnel.

Counts for Armed Forces and Foreign Armed Forces were then extracted on the basis of the Industry coding. Foreign Armed Forces were often incorrectly coded to UK Armed Forces. Adjustments for Foreign Armed Forces applied only to a few LAs where there was known to be high number of foreign Armed Forces.

No coding was changed as a result of this investigation, but people records identified as likely members of the Armed Forces were flagged in the Census Output database. This enabled more accurate information to be produced on counts of people believed to be members of the Armed Forces.

Census 2001 Review and Evaluation

Same Sex Couples

The validation process detected that there was an inflated number of same sex couples, both with and without children. The main reasons were:

- the form filler ticked the wrong sex for one person;
- duplicate instances of the same person;
- 'Partner' being imputed when the relationship matrix was incomplete or inconsistent;
- the One Number Census process adding a person into an existing household; and
- the Household Composition Algorithm deriving an incorrect family type.

A set of automated checks, including analysis of first names to check whether the wrong sex had been ticked, was carried out to identify people who fell into the above categories. Where couples could not be decided automatically the record's validity was decided manually. A few households containing more than one potential same sex couple were passed straight to manual checking.

The number of same sex couples counted in the Census in England and Wales was 88,322. Of these, 39,261 were found to be genuine (44.5 per cent). A further 18.0 per cent were imputed by the ONC process, 12.3 per cent were caused by the wrong sex being ticked, and 25.2 per cent were due to imputed relationship information. This information was collected to produce a Same Sex Couples database. To date, no data has been corrected or flagged on the Census Output Database.

Geographical Variations

In analysing the data various small anomalies were observed. In the majority of cases these were related to the particular demographics and industrial structure of the area.

For example, West Anglia had a high proportion of people born in USA due to United States Air Force bases in the EA. South Cheshire had large numbers of people working in Chemical and Car Manufacturing,

and Somerset had large proportions in Aircraft Manufacturing. In West Lancashire there was a high proportion of employees coded as Prison Service Workers, as four prisons were located in the EA.

Lessons Learnt

Whilst the Data Validation process was successfully completed the following lessons were learnt:

- Greater harmonisation of questions in successive censuses would allow for the easier generation and comparison of secondary data. Although much greater efforts were made than in the past to harmonise the 2001 Census questions, comparisons were still difficult. However, there will always be the constraint that most secondary data is based on relatively small samples, making comparisons of limited value because of the scale of sampling error.
- To involve at an early stage in the development of the Data Validation strategy, providers of secondary data and topic experts in ONS and other government departments.
- To trade off alternative validation strategies prior to the main census including where appropriate:
 - researching software tools to automate the process as much as possible;
 - match all secondary/comparison data checks to relevant data and assess the quality and reliability of the test, prior to specifying the limits;
 - estimate the number of Validation Checks to be carried out prior to development;
 - integrate the validation checks with other processing stages and produce a compliancy matrix to ensure coverage of checks and avoid duplication;
 - prototype the preferred test solution(s) software and hardware using realistic test specifications and gain metrics to measure how long each test takes;
 - provide sufficient time in the project plan for training of team members;
 - provide time in the plan for the integration of the required software and hardware;

Census 2001 Review and Evaluation

- reduce the repetitive nature of the tasks by automating checks where appropriate, so reducing the risk of human errors being made in the interpretation of the data;
 - establish criteria/ranking guidelines as to the importance of different types of errors and the remedial action that should be taken;
 - work closely with NISRA and GROS in developing new strategies to validate census data;
 - use supported software tools during the Validation period to avoid 'work rounds' and for the fast resolution of errors;
 - prototype the scanning technique to avoid the appearance of black lines on images; and
 - improve the collection of Armed Forces data.
- Although more could have been done to identify cost/resource/quality trade-offs in the development phase, there will always need to be further adjustments at the implementation phase. In particular, testing based on the dress rehearsal cannot be exhaustive because the volumes are so much smaller than for the Census itself.

Conclusion

The Data Validation strategy successfully evolved to meet the project's resource and time constraints.

Initially, it was found that reliance on secondary data and running a large number of AVCs was excessively time-consuming and diverted attention onto minor issues. The strategy was modified so as to target errors which could significantly degrade the quality of the Census data. Team members could then take informed decisions as to whether further investigation was justified when anomalies were found, leading to improved productivity and the project timescales being met.

In addition the revised strategy helped to resolve important quality issues including black lines, the discrepancy in the armed forces counts, the inflation in numbers of same sex couples and geographical variations in the data.

Census 2001 Review and Evaluation

Census Topics	Target Dates for Release
Legislation	Published
Non-Compliance (Executive Summary Only)	Published
Data Needs	Published
Geography	Published (Executive Summary)
Publicity	Published
Data Collection Development	Published
Data Collection Support	Published
Census Coverage Survey	Published
Processing	Published
Annex: Quality of Data Capture and Coding	Published
Downstream Processing	Published (Executive Summary)
Data Quality	
- Question non-response rates	Published
- Disclosure Control (Executive Summary only)	Published
- Data Validation (Executive Summary only)	Published
Edit & Imputation	Published
One Number Census	
- Quality Assurance	Published
- Lessons learnt (Executive Summary only)	Published
Output Policy	Published (Executive Summary)
Output Production	
- Part 1:Review of Output Released to date	Published (Executive Summary)
- Part 2:including Sample of Anonymised Records (SARs)/Origin Destination Matrices	Published
Census Access	Published
Programme Management	Published (Executive Summary)
Quality Report	Published
General Report	Published

Please note that the dates for release of individual evaluation reports noted above are target dates, and therefore subject to change. For the latest information please visit www.statistics.gov.uk/census2001/reviewevaluation.asp