2001 CENSUS: DATA VALIDATION

Project Objective

To ensure Census data can 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 caused by coding and scanning errors. There was a need to establish systematic checks and validate data at various stages throughout the processing of the 2001 Census.

Research was undertaken into how other countries approached Data Validation in order to identify best practice. It is common practice to compare census estimates with secondary data derived from valid surveys or previous censuses. A Data Validation Strategy was developed 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 Government District (LGD) level. The responsibilities of the team included specifying the tolerances for which the checks would be made, validating data, referring unexpected issues/errors to management for guidance and checking the solutions;

- Make use of Automatic Validation Checks (AVC) allowing checking to be more systematic and faster with a 'Pass', 'Fail' and 'Refer' classification. This approach would give the Data Validation team greater time to investigate discrepancies and possible errors;

- Cooperate with the Office for National Statistics (ONS) and General Register Office for Scotland (GROS) in validating their data by sharing developed software tools and knowledge. ONS 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 (i) interrogate and tabulate large volumes of data and then compare distributions with expected distributions, (ii) run queries down to record level where necessary and (iii) produce tailored reports to identify whether the Census data had 'Passed', 'Failed' or had been 'Referred';

- Undertake Validation at three critical stages of the project lifecycle, namely, at Load, after Edit and Imputation and after the data was ready for output; and

- Develop checks which highlighted inconsistencies in the data and, if necessary, to correct these using images of Census forms as reference.

Methodology

The initial validation methodology was based on the strategy of comparing Census data with secondary data gathered from previous Census and other surveys. Initially, the plan in Northern Ireland was to validate data at LGD level. 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 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 LGD 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 LGD.

Analysis

Once the DQMS report was complete for the LGD, 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, mostly because of the large number of LGDs in individual EAs;

- 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; and

- The relative level of the ‘Failures’ of the AVCs was unknown initially with minor anomalies being investigated at the expense of more significant errors.

Revision of Strategy

In recognition of these shortcomings, a revised strategy was implemented which tended to focus more on the ‘big picture’ and resulted in better targeting of available resources. This saw a change in emphasis from concentrating unduly on small anomalies which were unlikely to have a significant impact on the quality of data, to giving priority to the potentially more major errors.

In order to undertake this strategy, validation was carried out by EA, rather than at LGD level, with comparisons being made variable by variable.

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. For example;

- 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; and
- The EAs which had been validated using the original methodology provided comparative data that was used when validating data during subsequent stages of the process.

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 and was caused by lines of dust settling on the scanners causing black lines 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 qualifications, Irish language and activity last week.

In response to the problem, the contractor, Lockheed Martin, revised their cleaning procedures, which significantly reduced the occurrences of black lines, and revised their procedures for checking images. However, by the time these additional procedures were introduced, all NI Census forms had been scanned. The method used in Northern Ireland to resolve the problem differed to that in the rest of the UK; ONS used a modeling approach whereas the Northern Ireland method went back to images of the original forms. In Northern Ireland, the analysis was based on the idea that the build-up of dust would lead to the black lines appearing in clusters of processed forms. A separate data validation process checked that any ‘person’ in the database had valid information on at least 2 out of 4 key variables (name, date of birth, sex and marital status). This separate process had been designed to identify ‘rogue’ people created by, for example, people putting a line through the page to indicate no person. With space for 6 people in the form, black lines would lead to the clustered creation of ‘rogue’ people. When such clusters were identified, images for the ‘real’ people on the forms were examined to check that the black lines had not created false data entries. In total 12,634 corrections were made; 3290 amendments were

made for Irish Language, 2043 for Activity Last Week and 1887 for the Qualifications question.

Same Sex Couples

The relationship question allowed any two people to indicate that they considered themselves to be partners. Thus, although not designed specifically for this purpose, the question enabled the production of statistics on same-sex couples. It was recognised that these data would be of interest because of the lack of alternative data sources, and the data were subjected to a specific validation process. A number of cases of same-sex couples were created by the imputation process while (relatively few) others were found to be due to errors in the database.

A set of automated checks, including analysis of first names to check whether the wrong sex had been ticked, were carried out. 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 Northern Ireland was 643. Of these, 288 (45 per cent) consisted of census returns where people of the same sex had indicated on the Census form that they were partners. The information on these 288 couples was placed in a Same Sex Couples database that has been used to produce statistical output on same sex couples. It reflects the lower bound on the actual number of such couples.

Most of the others were plausible, but had been created by the imputation system. For example, some 11.4 per cent were imputed by the ONC process, and a further 29.8 per cent were created by imputation of missing relationship information. A small number of cases (0.31 per cent) are believed to be error caused by one person ticking the wrong sex box.

In summary, there are at least 288 same sex partner couples reported in the Census, and the Census imputation processes suggest that there are likely to be as many again.

Data Consistency

In analysing the data various inconsistencies were observed. In the majority of cases these were related to scanning errors and in a minority of cases errors due to incorrect coding.

For example, during the validation work a number of persons aged 0 to 15 were found to be married or re-married. Examination of the images of the original Census forms revealed that such instances were caused by errors in the

scanning of a person's date of birth. Such errors were amended to ensure the correct age was assigned.

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 NISRA and other Government departments.

- To trade off alternative validation strategies prior to the main Census including where appropriate:

    - researching software tools to automate and/or speed up 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 compliance 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;

- 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 ONS 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; and

- prototype the scanning technique to avoid the appearance of black lines on images.

- 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 potential 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 number of same sex couples and inconsistencies in the data.