Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

USING MATCHED CENSUS-SURVEY RECORDS TO EVALUATE THE QUALITY OF SURVEY DATA

Stephanie Freeth, Amanda White and Jean Martin[1]

ABSTRACT

Following the last three censuses in Britain, survey nonresponse on major government household surveys has been investigated by linking addresses sampled for surveys taking place around the time of the census to individual census records for the same addresses. This paper outlines the design of the 2001 British Census-linked Study of Survey Nonresponse. The study involves 10 surveys which vary significantly in design and response rates. The key feature of the study is the extensive use of auxiliary data and multilevel modelling to identify interviewer, household and area level effects.

KEY WORDS: Nonresponse; interviewer effects; household level effects; area level effects.

1. INTRODUCTION

The Office for National Statistics (ONS) is responsible for carrying out the decennial population census in England and Wales. Following the censuses in 1971, 1981 and 1991 investigations into survey nonresponse on major government household surveys were carried out by the Social Survey Division (SSD) of ONS which carries out many of the major government household surveys. In these census-linked studies of survey nonresponse, addresses sampled for surveys taking place around the time of the censuses were linked with individual census records for the same addresses. Since all the Census variables are available for both responding and nonresponding addresses, this provides a very powerful means of investigating the characteristics of nonrespondents, measuring nonresponse bias and evaluating methods of adjusting for the bias.

For the studies carried out following the 1971 and 1981 censuses, the actual matched datasets were not available to the methodologists investigating nonresponse, only specified aggregate tables, which limited the analysis to descriptive comparisons of the characteristics of respondents and nonrespondents, and measurement of bias in terms of census characteristics. However, in 1991 matched micro records were made available under strict confidentiality arrangements which allowed much more detailed statistical modelling to be undertaken to explore the interrelationship between variables which relate to nonresponse bias and to assess different methods of re-weighting data to compensate for the bias (Figure 1). Five surveys were included in the study in 1991, allowing comparison of results for surveys with very different designs. Key results are presented in Foster (1998).

The UK 2001 Census took place in April and in 2000, ONS began work on the 2001 Census-linked Study of Survey Nonresponse (CNR). This paper outlines the design of the CNR and describes the progress of the project to date.

[1] Stephanie Freeth, Amanda White and Jean Martin, Office for National Statistics, UK, 1 Drummond Gate, London, UK, SW1V 2QQ

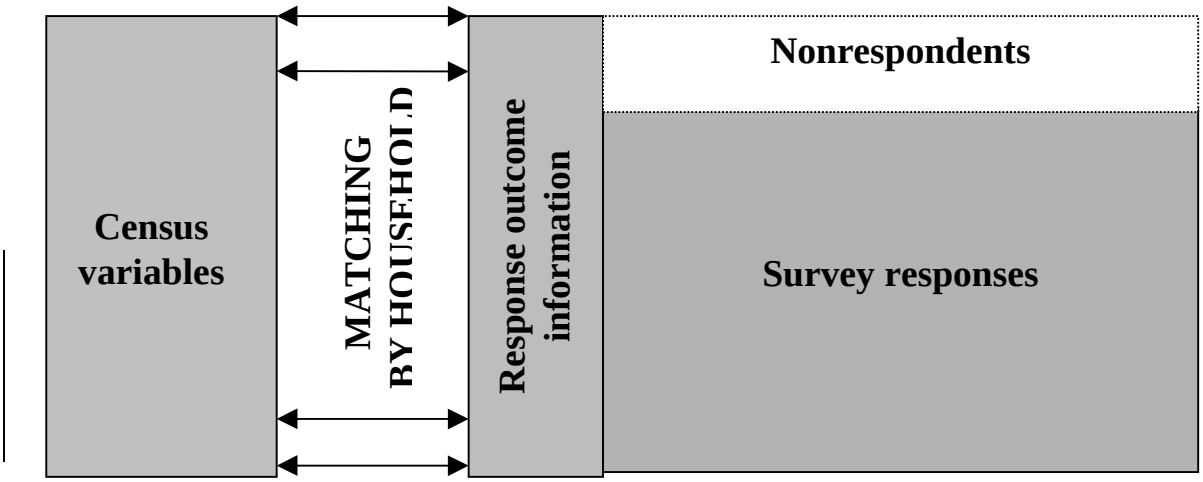

Figure 1 Design of matched dataset for 1991 Census-linked Study of Survey Nonresponse

2. DESIGN OF THE 2001 CENSUS-LINKED STUDY OF SURVEY NONRESPONSE

The scope of the 2001 CNR is wider compared with its predecessors. Ten surveys with varying survey designs and response rates (currently ranging between just under 60% to over 80%) are included. The 2001 CNR builds on the previous census-linked studies of survey nonresponse carried out by ONS and research conducted in the USA by Groves and Couper (1998). Groves and Couper have put forward conceptual models of the factors determining the likelihood of the interviewer making contact with a sampled household and the likelihood of the household agreeing to co-operate given contact. They list four broad categories of influence:

- area characteristics
- household characteristics
- survey design features
- interviewer characteristics

Each of these combines with the others to affect both likelihood of contact, the interaction between the household and the interviewer and hence the likelihood of co-operation given contact. Groves and Couper emphasise that it is not demographic characteristics per se which determine nonresponse; rather that people with certain characteristics are likely to lead lifestyles or hold attitudes which determine how easy they are to find at home or persuade to take part in a survey. Interviewers have a huge influence on response outcomes. The attitudes and strategies they bring to their work and their detailed behaviour at the household have been shown to be major determinants of response outcome. The CNR is designed to measure as many factors as possible which are likely to influence response outcomes in order to explore how area, household, interviewer and survey design characteristics interact to impact on nonresponse.

3. THE DATA TO BE USED

3.1 Overview of the Data

The data to be used in the CNR include census records and data on addresses/households sampled for surveys carried out around the time of the 2001 Census (which took place on 29 April 2001). The census data for households sampled for the participating surveys will be extracted and matched with the

corresponding survey data. The key feature of the 2001 study is the use of a significant amount of auxiliary data to supplement the information available from the census. These auxiliary data consist mainly of observational data about the selected addresses collected by the interviewers in the field. The study's dataset will contain the following information:

- Information about the areas sampled for the surveys
- Information about sampled households
- Survey design features
- Information about the interviewers
- Information about interviewer behaviour and outcome of visits to each sampled address

3.2 Area Information

Much of the information available about the areas sampled for the study will come from the census but information from other sources will also be included. The area variables have not been finalised but they are likely to include:

- population density
- whether urban or rural
- crime rate
- unemployment rate
- proportion of owner occupiers
- proportion of multi-occupied units

We will also have interviewer recorded information about the sampled addresses that might be predictive of ease of contact or gaining co-operation. For example, the state of repair of the buildings in the area, whether the area is mainly residential or mainly commercial. In addition, we will investigate various area classifications (ACORN, MOSIAC etc) that are currently available. Area level information can be linked to samples for any general population survey we carry out and is therefore available for routine nonresponse adjustment (and for sample stratification and analysis). The CNR will allow us to assess how well area level information compensates for nonresponse bias and how it may best be used.

3.3 Information About Households

The study's dataset will have all the variables included in the census. We will also have interviewers' observations of the characteristics of the selected households which might be predictive of ease of contact (eg presence of entry phone or other barriers to access) or gaining co-operation.

3.4 Survey Design Features

The surveys to be included in the study are all carried out by face-to-face interview but vary with respect to the following design features:

- whether information is required about all the individuals in the household or a selected adult
- whether all adults are to be interviewed in person or whether proxy information is allowed
- the response rules which determine the response rate
- the average length of interview
- the length of the field period
- the survey topics covered
- whether non interview data (eg diaries) are collected.

The dataset will include survey design information to enable the analysis to allow for survey design effects.

3.5 Interviewer Characteristics

In addition to information about sampled areas, households and survey design features, we will also include data relating to the characteristics of interviewers. These interviewer level data are from a survey of

interviewers ONS carried out in July 2001. This survey is similar to a survey carried out in 1998 as part of an international project (Martin and Beerten, 1999 and Hox, 1999). We have collected the following interviewer level data:

- socio-demographic characteristics
- length and nature of survey experience
- performance grade (based on response and other factors over a year)
- confidence in ability to gain response
- knowledge of techniques which encourage response
- reports of strategies used to persuade people to respond.

3.6 Interviewer Behaviour and Outcome of Calls

As with the earlier studies, we will include information about the visits interviewers make to each sampled address and the outcome of each call. We will also have other potentially useful information about the interaction between the interviewer and respondent when the interviewer introduced the survey at the doorstep. This doorstep information is recorded by the interviewer at the end of every call to an address. We have collected the following information:

- time of day and day of week of each call
- outcome of each call
- number of calls to make initial contact
- number of calls to complete interview after making contact
- doorstep interaction between the interviewer and the respondent, for example:
    - questions and comments of respondent
    - reasons given for not granting an interview.

4. OTHER DESIGN FEATURES

Each participating survey will have matched census-survey data on about 5,000 addresses. We plan to carry out a lot of analyses separately for each survey and a sample of 5,000 addresses will yield a large enough number of nonresponding households for survey level analyses. As in previous years, responding households will not be sub-sampled because this will have adverse effects on the statistical models to be developed. Addresses in Scotland and Wales will be included for surveys that cover these countries.
Ideally we want to match addresses which will be contacted as near to the census night (29 April 2001) as possible. This was carried out in previous years by including addresses selected for survey interviews in the months on either side of census night. In 2001, census data will be matched with survey data of addresses selected for interview from April. Matching addresses selected for interview before April is not feasible due to data compatibility problems linked to the adoption in April of the National Statistics Socio-economic Classification (NS-SEC) and the modification of a number of classificatory questions used on government surveys in Britain.

Figure 2 summarises the design of the 2001 CNR dataset.

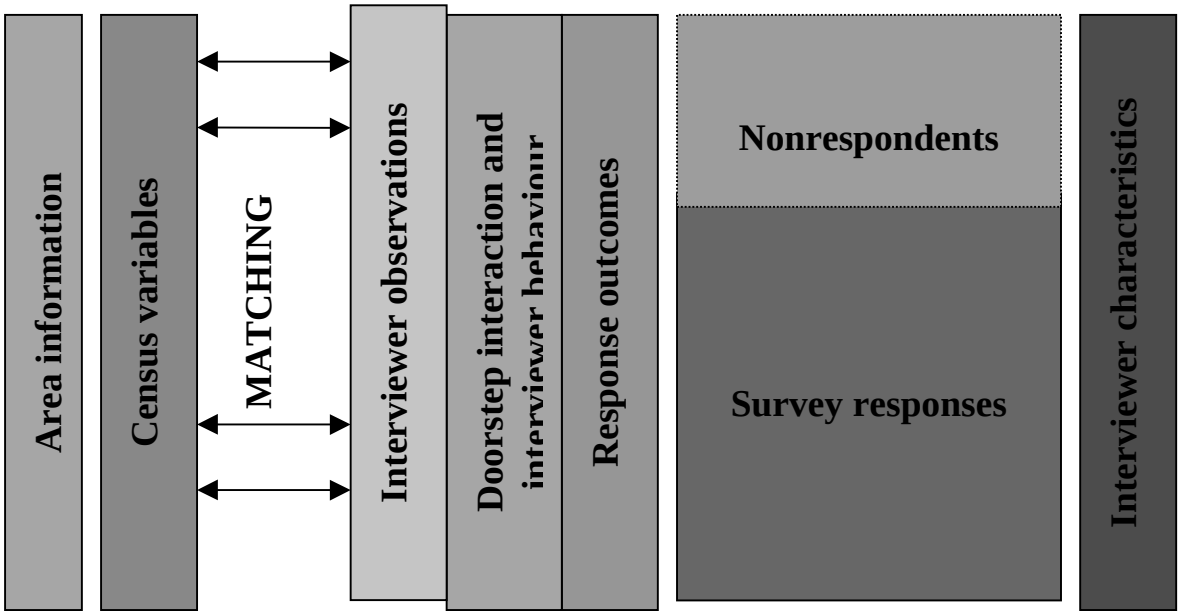

Figure 2 The design for the 2001 CNR dataset

5. ANALYSIS STRATEGY

Two aspects of nonresponse will be explored in the analysis: whether the interviewer made contact or not; and for each household where contact was made, whether the household co-operated or not. Account will be taken of the hierarchical structure of the dataset in the analysis and multilevel modelling will be used to distinguish effects at the different levels. Essentially we have three levels of information (Figure 3):

- interviewer – characteristics and attitudes
- assignment – survey design features, area characteristics
- household – interviewer observations, interviewer behaviour, household characteristics

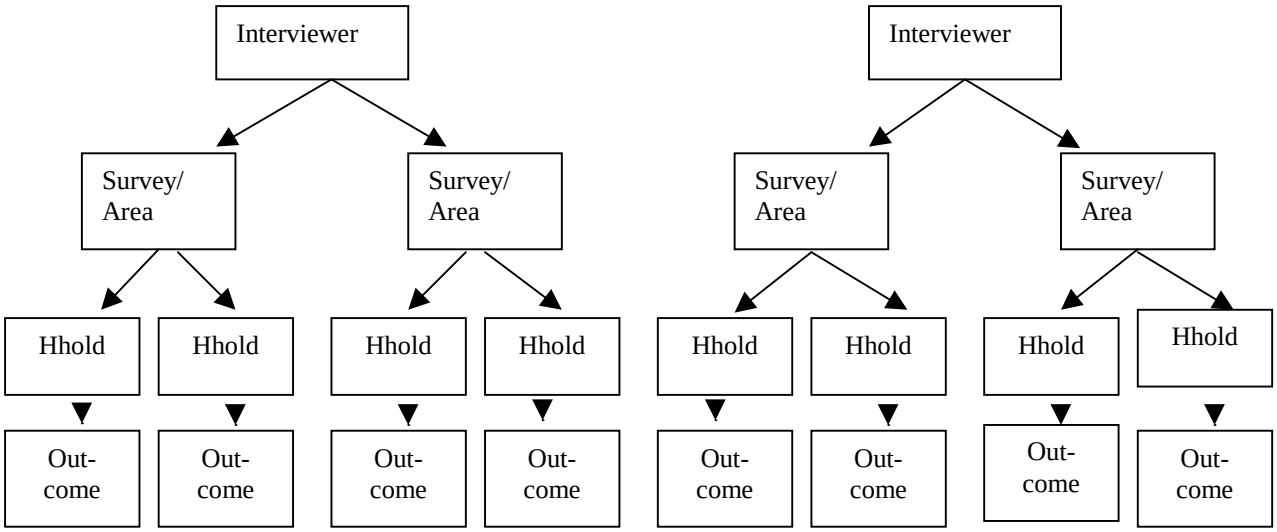

Figure 3 Data structure

Organisations sponsoring the surveys included in the study will be paying for their participation and will commission outputs for their particular survey. As with the 1991 study they will be offered a range of analyses from which to select. Apart from survey specific analyses we plan to use this rich dataset to examine influences on nonresponse across surveys. This follows the approach used for the analysis of the effect of interviewer characteristics on nonresponse carried out recently (Martin and Beerten, 1999).

Reports giving the results relating to a particular survey will be prepared for the organisations sponsoring each participating survey. Other methodological papers and reports comparing the results across surveys and drawing more general conclusions about the nature of nonresponse will also be produced.

The analysis options we will offer include:

- Descriptive comparisons of the characteristics of responding, non-contacted and refusing households.
- Measurement of nonresponse bias including summaries of over or under-representation of different subgroups in terms of census characteristics.
- Comparison of the above with results from previous years (if available) to identify changes over time.
- Logistic regression modelling to determine relative influence of different census variables on response outcomes.
- Multilevel modelling to determine the relative effects of area, household and, if available or appropriate, survey design and interviewer level variables on response outcomes.
- Development of weighting schemes based on the above analyses. Ideally these will incorporate information that could be collected routinely by interviewers as well as the information used in current nonresponse adjustment.

6. PROGRESS OF THE PROJECT TO DATE

Development work for the study started at the end of 2000. We have carried out a lot of work to design the procedures for collecting the interviewer observation data (for example, the characteristics/state of the area around the address, impediments to entry to the selected address and the interaction between the interviewer and respondent). The procedures were piloted in January and February 2001. The aims of the pilot were to:

- check that the data collection procedures operate satisfactorily when used simultaneously on a number of large continuous surveys
- assess the time interviewers need to collect the information to ensure that the data collection module can be accommodated into interviewers' assignments.
- obtain suggestions for improving the design to minimise burden on interviewers and to help develop training material for the main stage.

Ten interviewers working on a number of government household surveys carried out by ONS were involved in the pilot. The pilot interviewers provided feedback and advice on question wording and layout, and practical matters such as how to handle the material in the field. The key findings of the pilot were as follows:

- Interviewers could obtain the information required from observation or from their introductory conversation with the informant.
- On average, the interviewers took 10 minutes to record the observation data at each address and this task could be accommodated into interviewers' assignments
- The information required must be recorded as soon as practicable because the details could easily be forgotten. Interviewers working on the pilot recommended recording the information immediately after the contact with the household or on the same day if immediate recording was not possible.
- A compact and user-friendly paper interviewer observation form was needed because it was not always safe or convenient for interviewers to record the observation data directly into their laptop computer in their car.

We incorporated the findings of the pilot in re-designing the procedures and circulated the revised interviewer observation form to three of the original piloters and twelve other interviewers. This second round of consultation produced some suggested simplifications to the lay-out of the form which were incorporated into the final design.

The collection of the interviewer observation data began in May 2001 and data collection has been completed on all but one of the participating surveys. The survey of interviewers was completed in August and a response of 84% had been achieved. The census data will be extracted for matching in 2002 and we expect to begin analysis in 2003.

REFERENCES

Barton, J. (1999), "Effective Calling Strategies for Interviewers on Household Surveys", *Survey Methodology Bulletin*, 44, pp. 14-26.

Campanelli, P., Sturgis, P. and Purdon, S. (1997), *Can You Hear Me Knocking? An investigation into the impact of interviewers on survey response rates*, London: Social and Community Planning Research.

Foster, K. (1998), *Evaluating Non-response on Household Surveys*, GSS Methodology Series No. 8, London: Government Statistical Service.

Groves, R.M. and Couper, M. P. (1998), *Nonresponse in Household Interview Surveys*, New York: Wiley.

Hox, J. J. (1999), "The Influence of Interviewer Attitudes and Behaviour on Household Survey Nonresponse: an International Comparison", paper presented at the International Conference on Survey Nonresponse, Portland, Oregon.

Lynn, P., Clarke, P., Martin, J. and Sturgis, P. (1999), "The Effects of Extended Interviewer Efforts on Nonresponse Bias", paper presented at the International Conference on Survey Nonresponse, Portland, Oregon.

Martin, J. and Beerten, R. (1999), "The Effect of Interviewer Characteristics on Survey Response Rates", paper presented at the International Conference on Survey Nonresponse, Portland, Oregon.

Morton-Williams, J. (1993), *Interviewer Approaches*, Cambridge: Dartmouth.