**UK ONS Census-social survey linking project: Assessing non-response bias**

**Isobel Millward, Dr Katie O’Farrell; Social Statistics Hub, Methodology, UK Office for National Statistics**

The quality of surveys is often characterised in terms of the response rate. For social surveys carried out by the United Kingdom’s Office for National Statistics (ONS)[[1]](#footnote-1), where participation is voluntary, it is important that the factors that influence unit non-response are well understood. Unit non-response is a form of missing data in surveys and can be defined as failing to get a valid survey response from a sampled unit (in the case of household surveys, from an individual or a household unit). Many factors in a survey may lead to non-response: target population characteristics, survey content, interviewer training/experience, mode of data collection, the use of incentives, length of interview; and the complexity of these factors potentially interacting makes non-response inherently difficult to analyse. In addition, information on the characteristics and survey responses are usually only available for those responding to the survey, which means that if the characteristics or experiences of those who do not respond differ from those that do, the survey can be affected by non-response bias.

Social survey response rates have gradually declined over the past several decades and continue to do so. For example, response rates from the most recent performance report for the ONS Labour Force Survey[[2]](#footnote-2) (Quarter 1, 2020) show a 2.3 percentage point quarter on quarter drop across Great Britain (excluding imputed cases). ONS aims to reduce the risk of, and adjust estimates for, survey non-response through several mechanisms:

* We have previously undertaken studies into interviewer attitudes and characteristics, in order to study the role of sociological factors such as the response mechanism influencing non-response.
* Our data collection methods aim to reduce non-response; e.g., the use of incentives to increase cooperation rates.
* We apply statistical adjustments to the survey weights to adjust for survey non-response and in an effort to minimise non-response bias. For the purposes of this brief paper, we consider non-response to be categorised into failure to contact a sampled household, refusal to participate on behalf of the households and cases where the eligibility for the survey is unknown.
* For the last 5 decennial population censuses, a study (the Census non-response link study) has been carried out to examine non-response in social surveys by linking response and non-response to social surveys with the Census to give insight into the characteristics of those who respond or do not respond to social surveys.

We are currently planning a Census Non-Response Link study (CNRLS) to take place at the time of the next UK population Census in 2021. This paper will focus on the analysis plans for this forthcoming project and will propose approaches to assessing non-response bias

1. **Overview of the ONS Census Non-Response Link Study**

For ONS household surveys in general, samples are address-based. Individual addresses are sampled from the Royal Mail’s Postcode Address File (PAF), meaning in practice very limited information is known about the characteristics of non-responding households at sampled addresses and only information on the area is available. For the past five decades, the ONS has undertaken a Census Non-response Link Study (CNRLS), making use of the opportunity that arises from the decennial Census to overcome challenges in adjusting for non-response in ONS household surveys. The next CNRLS will go live in 2021 alongside the ONS Census for England and Wales. By linking sampled survey addresses from several household surveys collected around the time of Census with corresponding addresses within Census returns, analysts can make use of the rich auxiliary information from Census responses to understand the household characteristics influencing non-response in household surveys. This can be used to inform the efficacy of non-response models for social surveys.

1. **The CNRLS in 2021**

The aims of the CNRLS in 2021 are to link Census data to data collected from social surveys at the time of the Census. The exact scope of the social surveys that will be considered by the project is being reviewed, but the aim is to make the study as relevant as possible. Social surveys at ONS cover data collection on topics such as employment, education, health, wealth, income and expenditure. There are some similarities in the design and methods (for example, address-based sampling) but there are also differences (for example, some surveys are clustered, some are cross-sectional and others are longitudinal). The main purpose of the CNRLS in terms of understanding non-response is to assess the needs of a non-response adjustment to each social survey, recognising that the factors influencing non-response behaviour may differ across surveys. In addition, though, there is an opportunity to consider whether there are any common patterns across surveys, which can be used both to inform non-response adjustments, but also potentially to inform data collection approaches.

The sophistication of existing non-response models also differs between surveys. For example, the Living Costs and Food Survey, which collects information on household expenditure, includes a non-response weight that is built from Census data, the model for which was informed by an earlier CNRLS. However, some surveys, which have previously not been included in CNRLS, use simple area-based non-response models, built as logistic regression models from basic characteristics of a geographic area. This study therefore presents a valuable opportunity to test the effectiveness of such an approach.

At a high-level, the CNRLS design takes three stages:

1. Prototyping: the use of the CNRLS 2011 input sources and resultant linked dataset to prototype both analytical and data linkage methods for the 2021 study. This stage will be completed in early 2021.
2. Data linkage: Data linkage for the 2021 study will utilise both automated and clerical matching, using the linkage methods developed during the prototyping stage. We will link social survey response files collected during January-June 2021 with their sample files. These datasets will then be linked with Census returns as this data becomes available. The final, linked, quality assured datasets are scheduled for completion by end of 2021.
3. Analysis: Using the analytical methods developed during the prototyping stage, we will undertake non-response and mode effects analysis on the linked datasets. Results will be disseminated through a series of research articles released 2022/23 and incorporated as statistical adjustments into social survey estimates as/when appropriate.
4. **Challenges for CNRLS 2021**

The COVID-19 pandemic has had an impact on ONS social surveys, with the cessation of face-to-face interviewing in March 2020[[3]](#footnote-3). Since then, changes to mode have been introduced to surveys, moving surveys that usually are collected by interviewers in the field visiting sampled addresses to telephone and online modes. This presents both a challenge and an opportunity. Firstly, it presents a valuable chance for rich insight into online non-response. However, there is a challenge to maximise the value and relevance of the CNRLS in the face of changes to the mode of data collection. We would be interested to learn of other analysts approaches to undertaking non-response analysis and adjustments in the face of changing survey design.

1. **Research and development of CNRLS 2021 non-response analysis**

The approach to non-response analysis that we will use in 2021 has been informed by the previous CNRLS studies. We plan to use data from the 2011 CNRLS to prototype and test our proposed approach. This presents an opportunity to develop methods of analysis in advance of the period of the study, which, given the need to derive benefit from the study as soon as possible, will mean analysis work can start as soon as data are linked and ready to be used. The approach set out below is currently in development and we welcome feedback on the proposal.

In terms of the dataset we will use for prototyping, the 2011 CNRLS was made up of 3 household surveys: The General Lifestyle Survey (GLF), Living Costs and Food Survey (LCF) and Life Opportunities Survey (LOS). Although only the LCF is common between the 2011 and 2021 studies (the GLF and LOS no longer exist), there are sufficient similarities in overall design between the remaining household surveys linked in 2011 and 2021 to sensibly facilitate development and research on the 2011 dataset, that will reliably inform the methods chosen for the 2021 project.

The primary goal of prototyping our analysis approach is to enable us to choose a single method to be used in the 2021 study for identifying a set of characteristics that splits households into groups with similar response rates, in order to estimate response propensities. The secondary goal is to explore possibilities for expanding our definition of response propensity and, also, to explore the use of measures such as R-indicators as alternatives to a binary response outcome.

We will explore two different approaches for examining the association between household characteristics and survey response: parametric logistic regression models, and non-parametric classification and regression tree (CART) models. CART models are considered to have several advantages over logistic regression: many predictor variables can be included; interactions are identified by the model but do not have to be pre-specified as is the case in logistic regression; the groups are easily interpretable, and CART models consider missing data to be valid. It is worth noting though that non-parametric models have lower statistical power compared to parametric models.

Our analytical prototyping stage will be split into the following steps:

1. Build logistic regression propensity models:
   1. We believe that for several reasons we can improve upon the modelling approach used in 2011 and, therefore, think there is utility in exploring logistic regression methods:
      1. The 2011 team chose household level variables for inclusion in the analysis that had greater than an 80% match rate between the social survey and Census data at the linkage stage. However, methods for data linkage now commonly utilise false positive and negative rates as measures of accuracy, rather than flat linkage rates. These measures will be utilised in our 2021 project and will give a more robust indicator of data quality, which may help to inform variable selection.
      2. Further, in terms of variable selection, a more structured approach to identifying relevant descriptive variables can be taken. This could include identifying variables for which response rates differ most dramatically, and those with potential interactions, and reducing the list through backwards stepwise regression before undertaking logistic regression modelling (see Carlson and Williams, 2001).
      3. Upon review, the methods for the current non-response adjustment applied in the LCF for example, are more akin to a hybrid weighting class/logistic regression approach rather than a pure logistic regression approach. Carlson and Williams (2001) note that weighting class adjustments are suitable when there is limited auxiliary data available for the weighting classes, which should not be the case in the CNRLS. A more typical logistic regression approach may therefore yield different results in terms of examining non-response bias.
   2. With these updated models, we will explore possible non-response bias for where we have outcome (Census) variables for both responding and non-responding households. We will assess whether any interactions of household characteristics associated with response propensity are also associated with outcome variables. This is in line with the stochastic view of non-response bias, whereby non-response bias is a function of how correlated the survey variable is to the propensity to be measured in the target population.
2. Build CART models:
   1. Previous research has utilised both classification tree models (see McCarthy, Jacob, and McCracken, 2010, for an example) and regression tree models (see Phipps and Toth, 2012 for an example) whereby iterative partitioning is used to describe the association between sample characteristics (independent variables) and propensity to respond (dependent variable; see Loh, 2011, for an overview of commonly used algorithms). We are interested in exploring both these machine learning methods, noting that classification trees are designed for categorical dependent variables (suitable for response/non-response), whereas regression trees are designed for continuous dependent variables (potentially more suitable for alternative measures of response such as R-indicators; see 4b below).
   2. We are also interested in taking advantage of our multiple-survey linking methodology to understand the impact of incorporating sample design information into models; whether that is by building separate models based on survey collection mode (telephone, online, or hybrid) or including mode as a categorical predictor variable in a combined model.
   3. Our work here will consider the most appropriate cross-validation way to test the accuracy of these models and establish the decision-making criteria for choosing one approach for our live 2021 project.
   4. As above in 1b, with these models, we will explore possible non-response bias for where we have outcome (Census) variables for both responding and non-responding households.
3. Evaluate performance of chosen CART model relative to logistic regression. This needs to consider:
   1. The models’ ability to identify and interpret interaction effects;
   2. A comparison of results in terms of identifying the characteristics of the groups least likely to respond and the potential impact on non-response bias
4. We are also interested in different approaches to the definition and measurement of propensities and response rates:
   1. the inclusion of other variables known for all sampled units when calculating the response propensity. Attempts to capture the effort or activity involved in data collection activities – for example, number of call attempts, the use of incentives, the modes of data collection – and explicitly include these when calculating response propensity are suggested by some researchers to better define the quantity being estimated (see Brick, 2013).
   2. exploring the use of different measures that are more informative of non-response bias than response rates, given the strength of our auxiliary data. These measures include the use of R-indicators, balance indicators, and the proportion of missing information as alternatives to response rates.

The prototype of the analysis described above will be used to develop a detailed analysis plan for the CNRLS in 2021. With that in mind, we would welcome feedback on the approach, including any alternative methodologies that could be of value.

1. **Specific questions for discussion:**

There are several areas we would welcome discussion on:

* Experience of using CART models as compared to logistic regression. What are the best methods for evaluating the performance of these different approaches?
* Experience of using measures that are more predictive of non-response bias. What experience does the audience have with different approaches to calculating propensities?
* How are other institutes dealing with design changes and accommodating non-response adjustments during Covid-19 which in many cases has resulted in altered sampling methodologies?

We are also interested in engaging with any international working groups, particularly those concerned with alternative approaches to understanding non-response bias, and/or producing quality non-response adjustments for surveys impacted by recent changes to sampling methodologies.

*Disclaimer: The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of the United Kingdom’s Office for National Statistics.*

1. **References**

Brick, J. Michael. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics, 29:3,* 329-353

Carlson, Barbara Lepidus, and Williams, Stephen. (2001). A comparison of two methods to adjust weights for nonresponse: propensity modelling and weighting class adjustments. Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

Loh, Wei-Yin. (2011) Classification and regression trees. *WIREs Data Mining Knowl Discov, 1,* 14-23

McCarthy, J.S., Jacob, T. and McCracken, A. (2010). Modelling Nonresponse in National Agricultural Statistics Service Surveys Using Classification Trees. Research and Development Division Research Report Number RDD-10-05, US Department of Agriculture, National Agricultural Statistics Service.

Phipps, P. and Toth, D. (2012). Analysing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics, 6:22,* 772-794

1. The UK’s National Statistical Institute [↑](#footnote-ref-1)
2. <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyperformanceandqualitymonitoringreports/labourforcesurveyperformanceandqualitymonitoringreportjanuarytomarch2020#response-rates> [↑](#footnote-ref-2)
3. <https://www.ons.gov.uk/news/statementsandletters/ensuringthebestpossibleinformationduringcovid19throughsafedatacollection> [↑](#footnote-ref-3)