Data quality

Strand organizer: Phil Humby (Office for National Statistics)

Data quality – Monday 9 September, 1.30pm

Data quality in mixed-mode mixed-device general population UK social surveys: Evidence from Understanding Society Wave 8

Olga Maslovskaya, Gabriele Durrant, Peter W.F. Smith

University of Southampton

We live in a digital age with a high level of use of technologies. Surveys have started adopting technologies including smartphones for data collection. There is a move towards online data collection in the UK, including an ambition to collect 75% of household responses online in the UK 2021 Census. However, more evidence is needed to demonstrate that the online data collection will work in the UK and to understand how to make it work effectively. This paper uses the first available in the UK large-scale mixed-mode mixed-device survey, Understanding Society Wave 8, where 40% of the sample were assigned to the online mode of data collection. It will allow comparison of data quality between different devices within the online mode. This analysis is very timely and will fill this gap in knowledge. We use the main survey of Understanding Society Wave 8. Descriptive analysis and then linear, logistic or multinomial logistic regressions are used, depending on the outcome variables, to study data quality indicators associated with different devices used by respondents in the online part of the survey. The following data quality indicators will be assessed: break-off rates, item non-response, response-style indicators, response latencies, and consent to data linkage. Comparisons with earlier results from the Understanding Society Innovation Panel and with results from other countries will be drawn.

Email: om206@soton.ac.uk

Determining where people are: An analysis of geocoding

Julian Buxton, Phil Humby

Office for National Statistics

When people respond to a survey or are recorded on an administrative system, they log address information, but do we record them where they actually are? We wanted to assess the variation that could be generated from a simple change to the geocoding, specifically the version of geography applied to a dataset through the geocoding method. The basis for the application of a version of geography on some administrative datasets at ONS has tended to be the latest one available when the data are received rather than by scientific determination. We wanted evidence to determine the validity of this approach. We took a comprehensive administrative dataset and applied a single method with five versions of geography to gauge the likely impact this might have across the population. We assessed the impact at four levels of geography – local authority, MSOA, LSOA and OA – by looking at the change in total population for a given area. While the levels of geocoding were high throughout, more people were geocoded as time went on, suggesting that later versions better captured housing developments. The results showed there was variation, with some people being allocated to different areas while address information was unchanged. There were some interesting quirks in the geographic allocation that were corrected over time. The results also showed that there were more reallocations at lower levels of geography. The

principle conclusion was that the existing basis for determining the version of geography used in geocoding was sufficient for production of population statistics.

Email: philip.humby@ons.gov.uk

But if it's a census, why do you need weights?

Viktor Racinskij, Ceejay Hammond, Ercilia Dini, Danielle Burke, Owen Abbott, Gareth Powell.

Census and Population Statistics Methodology, Office for National Statistics

Every effort is made to ensure everyone is counted in a census. However, no census is perfect, and some people are missed or counted twice. This imprecision does not usually occur uniformly across all geographical areas or across other subgroups of the population such as age and sex groups. The measurement of small populations, one of the key reasons for carrying out a census, is becoming increasingly difficult. In terms of resource allocation, this is a big issue since the populations that are missed can be those which attract higher levels of funding. Therefore, without any adjustment, the allocations based upon the census would result in monies being wrongly allocated. It is therefore traditional that census undercount and overcount is measured and the outcome disseminated to users.

This paper outlines the proposed strategy for the 2021 UK Census. While the methodology is applicable to the UK, it is expected that there will be slight differences between countries to reflect local circumstances. As in the 2011 Census, a Post Enumeration Survey (PES) will be used. Building on the Dual System Estimation approach used previously, we intend to use logistic or mixed effects logistic regression models to produce weights. As the 2021 Census will be primarily completed online, we are also developing our index of Hard-to-Count areas to take digital access into account.

Email: gareth.powell@ons.gov.uk