Statistical Design for Census 2031 – MARP Discussion paper

Statistical Design Team, Census Taskforce, ONS.

Introduction

This paper sets out how ONS are approaching the statistical design and methodological aspects for Census 2031. It is intended to give MARP early sight of some of the thinking and to provide an opportunity to feed ideas early into the Census Taskforce.

The taskforce aim is to set Census 2031 up for success by backing the right ideas and setting the foundation for the development of the end-to-end design. Given the relatively late start to planning for 2031, not all ideas can be fully explored within the time available and carrying untested options into the operation is too risky. The objective is to be deliberate in selecting the right innovations to pursue that balances risk against potential benefits.

Our approach is to surface as many ideas, potential innovations and improvements as possible during the taskforce phase, so that we can narrow down the options in advance of setting down the critical path through to 2031. The driver is landing a well-balanced and realistic business case with Treasury by the end of 2025 which gives confidence in delivery – and provides a balance of risk and innovation.

The focus during the taskforce period is to evaluate innovations and improvements that have a significant impact on costs, so while improvements to (say) statistical processes like Edit and Imputation will be explored, these would not be within the time frame of the Taskforce and will be planned for subsequent years.

One key consideration is that whilst the primary objective is to deliver Census statistics that meet user needs, a secondary objective is for the census to provide a stepping stone to a future system which increasingly uses more administrative data. This second objective may result in some choices which enable better linkage or data collection in the future, but these will be carefully evaluated as they may risk the primary objective. An example could be the inclusion of a question which asks about administrative identifiers such as national insurance number.

MARP are asked to:

- Comment on the innovations in the Statistical Design that have been identified
- Suggest any other innovations or improvements that ONS should consider
- Note that we will return with a paper setting out the innovations we are taking forward and those that we are not

Statistical Design

A small team has been established to develop the basic statistical design for Census 2031 and evaluate potential improvement and innovations.

The approach taken is to build on the 2021 statistical design framework published in 2020. We have undertaken a rapid assessment of the base features of the design to explore whether some of the fundamentals could be improved over the 2021 model. In doing so, we have tried to identify assumptions and challenge those to ensure they are robust.

The following are our initial proposals, that we have considered so far:

a. The usually resident population should be used as a population base for collection and outputs in Census 2031
b. We will adopt the same quality goals as in 2021 and will aim for first outputs within a similar timeframe. Key metric requirements for the operations (e.g. response rates) will need to be developed.
c. An address frame should be used for 2031 Census, using AddressBase as the underpinning product.
d. We recommend that we invite the whole population to take part in the census, but we should explore whether there are specific well-defined populations where we can use administrative data.
e. Continue with at least the level of targeted follow-up as in 2021 (area-based) and investigate the opportunities to use administrative data and Machine Learning.
f. A fully coverage adjusted output database is required, and the methodology and system needs developing and including in the critical path planning.
g. We will measure coverage using statistical techniques. Research into whether we need a coverage survey is needed.
h. We will edit and impute the census data to deal with inconsistencies and missing values. We should determine whether CANCEIS is an option.
i. We will assume we will do the same SDC methods as 2021, unless any new techniques emerge, or if we need to do something additional due to the use of administrative data.
j. We do not release census counts as well as census estimates. We could do an early release of high level (rounded) estimates before processing is finished.
k. We will explore the use of administrative data across the whole Census process, from quality checking of address lists to potential coverage adjustment for estimation.

Statistical Design Innovations

Some of the big statistical design innovations that have been identified are set out below, with some initial thinking around their feasibility and risk from a quality perspective. Many of these are essentially aspects of the field operation, because that is where the most benefit and risks are:

Better frame, especially for Communal Establishments

The address register used in Census 2021 had some issues with classifications of Communal Establishments which led to issues with the collection. Additional priority to build a more robust register utilizing more sources (and early engagement with CE managers) is something we feel is key innovation worth pursuing. Statistics Canada use their survey telephone interviewers to undertake a CATI with managers well in advance of Census to improve the frame.

Better targeting of follow-up

Administrative data was used in Census 2021 to help target the field allocation and follow-up through the response chasing algorithm. This information was at an area level. One option for innovation is to develop this further using the various sources of information we now have at address level, including information about the likely household size and composition. This could include predictive models for vacant properties (the US and Canada are doing this), or even a predictive mode for likely mode or timing of response. These models could be retrained as responses are received, and the predictions used to drive follow-up actions (e.g. publicity, reminder letters, SMS reminder, field visit).

Use of administrative data for measuring coverage

We are assuming that we will need to measure and adjust for coverage errors. Whether administrative data can do this entirely (and not require a Census Coverage Survey) will need to be decided by the end of the Taskforce. Our initial view is that this is unlikely and it is probably too risky to rule out a CCS. However, individual level administrative data could be used in some capacity as it has done in both Northern Ireland and Scotland, alongside a CCS, which requires research.

Pre-population

One idea is to provide a mechanism for people to see what data is held for them (on administrative data or perhaps previous censuses) to reduce burden in completing the questionnaire. This would be challenging from a privacy perspective as some form of robust authentication would be required. There may also be quality risks if people don’t make corrections, and reputational risks if those corrections cannot be passed back to data owners. We think this innovation is unlikely.

Replacement of census variables and adding new content

In Census 2021, user needs for information on number of rooms was met using Valuation Office Agency data. The intention was to also provide an income variable

using PAYE and Tax information, but this was de-scoped. There is the opportunity for 2031 to evaluate which variables could be wholly replaced, and which variables could be added to the Census data in some way. Income would be extremely valuable for users given the high demand for it, and we see this as a priority for development.

Elimination of paper

There are operational challenges associated with scanning and processing large volumes of paper, and costs of delivering paper questionnaires are significant. Eliminating, or reducing paper to a very small proportion could offer significant savings and efficiencies. However, this would have to be balanced against the risk. Our expectation is that we could reduce paper as much as possible but are likely to need some paper questionnaires as the costs of alternative provisions are likely to exceed savings by not posting and processing paper. There is also a risk to response through not offering an easy route to complete on paper for some parts of the population.

Do we have a single set of questions for the whole population?

Do we ask the same set of questions to the whole population, or can we adopt a modular approach for content on the Census questionnaire?

Modularised content could be based on geography (e.g. region) - where the user needs are different for different regions (regions could be mayoral areas). There is precedent for this where a Welsh language question was included on questionnaires in Wales only in 2021 (Q17 - ‘this question is intentionally blank’). Modularised content can also be based on individual characteristics - different modules are activated dependent on respondent responses e.g. age or language or activity. This is very much like regular questionnaire routing. Lastly, modularised content could be based on a random selection (e.g. 20%) of the frame, providing a sample which should be sufficient to provide LA level estimates.

These could be delivered through different questionnaires or routing in an online/telephone questionnaire. Would be likely impossible for paper, which might have to be just core questions with no modular add-ons.

Whilst technically feasible, this depends on the strength of user needs and the level of complexity and risk this adds.

Summary

The potential innovations outlined in this paper is not exhaustive and the Taskforce is keen to expose all ideas and potential options. Once we have settled on some base design principles and have scoped out the innovations and improvements that need to be tested as part of a 2027 Test, further innovation development becomes risky

and ultimately will drive up cost. Our approach is to explore these potential innovations early and converge on the likely design during the latter parts of 2025 leading up to the business case.

The statistical design is a key part of this, and we are making good progress on setting out some of the basics and ensuring that any innovation is focused on the primary objectives of producing statistics that meet user needs.