United Nations

ECE/CES/GE.41/2022/12

Economic and Social Council

Distr.: General
12 July 2022

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-fourth Meeting
Geneva, 21−23 September 2022
Item 2 of the provisional agenda
Lessons learned from censuses of the 2020 round

Administrative and field paradata sources for quality assurance and contingency approaches

Note by Office for National Statistics, United Kingdom*

Summary

Before the 2021 census in England and Wales we set out our proposals for uses of administrative data in our design, and the additional management information data we were getting from our online response and technological innovations in field follow-up. We were able to use the gathered data in various ways, including aspects that we hadn't planned for, that arose out of the COVID-19 pandemic.

This paper sets out the reality of how we have made use of various sources, including checking for any biases in our coverage estimation approach, and assessing and using in contingency approaches across a range of areas.

* Prepared by Cal Ghee.

Note: The designations employed in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

I. Introduction

1. Administrative and other alternative sources have played a key role in aspects of previous censuses. They were used extensively in the operational planning and quality assurance of the 2011 Census. For 2021, we set out in advance what additional uses we were exploring across the whole census operation (UNECE, 2018).

2. The United Kingdom (UK) Census is made up of three separate operations: Scotland and Northern Ireland developed their own design based on the most appropriate data and methods for their areas. This document only covers the work of ONS on the census for England and Wales.

II. Background

3. The UK censuses primarily use data collected from self-response from every person and household. Each address received an invitation letter containing a unique online access code, or a paper questionnaire with the online code, with instructions to submit their information around Census Day: online, by post or, if necessary, by telephone. Not all households do this straight away, so the collection operation involved a series of non-response follow-up initiatives, including:

(a) Ongoing media messaging;

(b) Reminder letters;

(c) Community liaison and events;

(d) Reminder visits by field officers.

4. Table 1 below summarizes the potential uses of administrative data we set out before the live operation. See also UNECE (2018).

Table 1
Potential uses of alternative data in the three stages of the end-to-end 2021 Census operation

Stage	*Use*	*Comments*
Prepare and collect	Creation of census address frame	Sources feed into validation of addresses, addition of new addresses
	Predicted response patterns, improving efficiency of field operation	Understanding the local population characteristics, and their non-response tendencies
Process and analyse	Adding to indexes and classifications	New and changing categories found in administrative sources can be added to indexes and classifications in advance, leading to higher rate of autocoding of variables
	Cleaning and editing	Comparison of responses against alternative sources
	Coverage bias adjustments	See sections below
	Placeholders in record imputation	Imputed households are placed in a geographic location from which they are likely to have been missed. The intention was to use census operation and administrative data to inform decisions about where this is appropriate

Stage	*Use*	*Comments*
	Quality assurance/validation	A range of sources has been used to validate the final census estimates of people, households and their characteristics
	Adjust for collected data in communal establishments (also known as collective dwellings or group quarters)	We calculated the shortfall in responses in communal establishments by comparing response levels by age and sex against administrative data sources
Outputs	Replace previously collected variables or creation of new variables	We have replaced the question on number of rooms with data already available in a high quality administrative source. We will produce integrated outputs on income, floor space and other topics.

5. More information on the stages above is available in our published statistical design (ONS, 2020). Further information is given below on the uses in coverage bias adjustments and the underestimation of students.

III. Sources

6. We used two main types of data in our design and contingency planning: administrative data and paradata.

A. Administrative data

7. This refers to information collected primarily for administrative reasons (meaning not initially for statistics or research). This type of data is collected by Government departments and other organizations for uses such as registration, transactions and record-keeping, usually as a by-product of delivering a service. Administrative data are often used for operational purposes, and their statistical use is usually secondary.

8. Examples include:

(a) Council Tax − information on residential addresses liable to pay for local services. This source is generally considered good quality as a local authority relies on this to fund services, and so maintains the administration closely, and a resident who has moved out will prefer to stop paying as soon as they move. However, definitions are not always consistent with the census definition of a private main residential household. Also, rules vary across local authorities, for instance whether discounts or exemptions apply for second homes;

(b) Health registrations − basic demographic data on every person registered with the National Health Service. This source has high coverage of the total population, with small undercoverage of people who only have private health care, or for new migrants. Dependent on people updating their details when they move, there is therefore evidence of under- and over-coverage at local area level, especially in areas of high population churn. For some groups (especially young adult males, who rarely need to interact with the health services), records can be out of date for a few years;

(c) Utilities − information on most addresses with active gas or electricity supply. This was explored to give information on addresses or areas with low activity;

(d) Work and pensions − data on people's interactions across a breadth of work and pensions systems;

(e) Student hall and boarding school survey − administrative data collected direct from university halls of residence accommodation officers, on the number of students with contracts to live in their accommodation for the census period;

(f) Birth registrations − data collected under legislation for statistical use;

(g) School census − information on all pupils attending state-funded schools in England and Wales. Data quality is high because schools are dependent on the accuracy of the data to get funding. However, it does not cover pupils in private education or those who are home-schooled;

(h) Electoral Register − names and addresses of every adult registered to vote in the UK. It is a legal requirement to register to vote if you’re asked to do so and you meet the conditions for registering. However, in practice the register tends to be most current around the time of key elections, and can get out of date between these.

9. Other sources were used in different parts of the design. For instance, data on the numbers of members of the Armed Forces were obtained from the UK Ministry of Defence and United States Armed Forces; the UK Home Office and Ministry of Justice provided data on people in immigration centres and prisons. For care homes, we used the health data mentioned above, and also care home Capacity Tracker data which were kept at a high quality especially during the COVID-19 pandemic for monitoring vaccination rates.

B. Paradata

10. These are data that describe the process by which the data were collected, normally through a survey. This includes the management information collected during the census operation. Examples of paradata topics include:

(a) The times of day that responses were submitted;

(b) Time taken to complete the questionnaire;

(c) Number of attempts to complete the questionnaire;

(d) How many times field officers called at non-responding addresses, day of the week and time of day, how many times they made contact and whether a response was subsequently received;

(e) Mode of communication (such as phone, web, email, or in person).

11. Therefore, there are paradata about each observation in the survey. These attributes affect the costs and management of a survey, the findings of a survey, evaluations of field officers and inferences someone might make about non-respondents.

12. For the first time in the 2021 Census we had access to 100 per cent of our collection management information on form submissions and on the field operation.

IV. Uses of the data beyond the planned design

13. The following sections summarize some of the uses we made of administrative sources and paradata, that were adaptations to the planned design. The focus below is on areas where we needed to make changes to the design, in order to meet our end-user needs for high quality census outputs:

(a) Students;

(b) Independent estimate of occupied households;

(c) National comparison of key age groups;

(d) Other age groups;

(e) Extra validation: houses of multiple occupation.

A. Students

14. A large proportion of UK students study away from their family home. On the census form, however, we know that a lot of parents will include children studying elsewhere. Therefore, we explicitly ask them to fill in these cases, but specify where their term-time address is. If not on the form they are filling in, that student will be filtered out of any further questions. When a respondent answers saying this is their term-time address, they get the full suite of questions to answer. This is effectively additional data collected not for the main purposes of the final census estimates, but for ensuring that we capture students only once, and at the address where they live most of the year.

15. We did extra communications and marketing of what was required of students for their census submissions, including instructions to fill in for their term-time address if they still had a contract to live there, but may have been temporarily located back at their parents' address due to the COVID-19 pandemic. See (ONS 2021) for more information on extra measures we took around the enumeration of students at the time of collection.

16. We could tell from the return data that we were not getting all the student responses we were expecting. However, we were getting responses on parents' forms, saying their child had a term-time address elsewhere. We were therefore able to use this data to place those students at their correct term-time address (having done a matching exercise to establish those who had already responded at that address).

17. This was a contingency that we put in place because of the impact of the pandemic. Unfortunately, due to the legislation process it was too late to change the questionnaire to stop filtering out the students at their out-of-term address from answering all but the key demographic questions, but still a bonus to be able to get additional use out of the basic information from data that was initially designed to prevent over-coverage.

18. Alongside this, we also recognized that the temporary relocation of students during the pandemic meant that our enumeration of halls of residence, as designed, would struggle to capture the agreed definition of who should have been enumerated at those addresses. The main data source on Higher Education students, from the Higher Education Statistics Agency (HESA) would not usually deliver the required period's data until later in the year.

19. Therefore, we set up a new survey which went out to universities and private providers, asking them for information from their own administrative data on the numbers by age and sex of students who still had a contract to live in their accommodation around census day, even if they weren't actually there.

20. This survey had a 92 per cent response rate, so we were able to use that information to estimate the true population of those halls, calculating the shortfall after taking responses and the above 'copied' records into account. For establishments that were not able to provide us with that data, we used other available information from the collection operation, number of bedspaces on our initial census frame, and borrowed strength from what we knew about undercoverage from the responses we did get, to estimate their shortfall too.

B. Alternative household estimate

21. The published UK census results are an estimate of the population, taking non-response into account using a Dual System Estimation (DSE) approach for assessing and correcting for under- and over-coverage. This is described in our statistical design (ONS, 2020), and a simple guide to DSE is set out in ONS (2011). DSE relies on two data sources: the collected census data, and the Census Coverage Survey: an independently collected post-enumeration survey covering approximately 1.5 per cent of households.

22. DSE methodology relies on a number of assumptions about the two sources, including assumptions about independence (i.e. that an individual's likelihood of responding to the CCS was not influenced by whether they responded to the census) and homogeneity of capture probabilities (i.e. that the population to which the DSE was applied has similar response patterns in both the census and CCS). If these assumptions do not hold, estimates

would tend to be negatively biased: i.e. resulting in estimates that are lower than they should be.

23. A bias could occur if, for instance, there was a tendency for households to refuse to respond to the census and the CCS, or if households that had responded to the census would then be more likely to respond to the CCS. We assessed for any bias that would be caused by these assumptions not holding true, by comparing household DSEs against an independent estimate of occupied households (called the Alternative Household Estimate (AHE)). The AHE was created using information from responses, field observations and other census collection information, and a range of administrative sources available for households.

24. The AHE was created from a range of alternative address-based sources:

(a) The census address frame;

(b) 100 per cent of census response management and field visit paradata;

(c) Local council tax data;

(d) Health registrations;

(e) School Census;

(f) Utilities;

(g) The 2011 Census.

25. The approach for creating the AHE used a basic model, using the relationship between the responses where people confirmed the presence or absence of residents, and what our field officers observed. We created indicators of either occupancy or vacancy from the administrative data (for instance: if Council Tax data indicated that an address was likely to be vacant because it was receiving an empty home discount; or if an address had a child registered on the School Census, it was likely to be occupied by usual residents). Each address was allocated to a group, and the occupancy rates of the responding addresses in those groups were applied to the non-responding addresses in that group.

26. We also had some deterministic rules based on field observations:

(a) If an address had received five or more field visits, by at least two different field officers, and each visit recorded that the address was vacant, this was set to vacant;

(b) If the field officer got their information about the address from a trusted third party (the householder themselves, a neighbour or a gatekeeper/doorman), the address was set to the field observation (i.e. occupied or vacant).

27. The intention was to use the AHE to assess for any between-household bias. In fact, it was used for other issues we found in the data:

(a) Bias in estimation: only one Local Authority was identified as having estimation bias for numbers of school children and the number of occupied households. This area had the AHE applied, as was originally envisaged in such circumstances;

(b) Higher levels of uncertainty (variance) in estimates: the CCS performed better in some areas than others. In areas where the CCS did least well, the estimates had a higher level of uncertainty indicated by estimates of variance from provisional confidence intervals. Outlier areas where the estimated number of households was lower than the AHE were calibrated up to the AHE value;

(c) Person response rate higher than household response rate: the AHE demonstrated that in some areas where there were apparently more households missing than persons missing, the household estimates were more robust. In these areas, the person response rate was calibrated to the household response rate.

C. Low estimated number of babies and very young children

28. Validation against health data and birth registration data demonstrated that the numbers of babies and very young children (0−2 years) were too low. We estimated the total population at each age, derived from data on births and deaths, and calibrated each Local Authority estimate to that total for ages 0, 1 and 2. To avoid overestimation in areas, health data were used as an upper limit.

D. Low estimated number of 3- to 15-year-olds

29. The Welsh Government maintains a source of data on all school-aged children: a combination of School Census for state-funded schools and other sources for children in private education and those who are home-schooled. This source identified that the number of 3−15 year old children was low in Welsh local authorities, so we were able to calibrate estimates to this source. Analysis of estimates for English regions demonstrated that the North East region was a similar outlier, so the adjustment factors calculated for Wales were also applied to each area in the North East region.

E. Assessing coverage in unrelated households

30. Houses of Multiple Occupation (HMOs) are particularly difficult to enumerate, because by their nature they contain people unrelated to each other. The risk is that one person may submit a census response just for themselves, and that addresses would then be removed from any further follow-up, but may still be missing all the other residents.

31. We mitigated this risk as far as we could in the creation of the address frame, by including HMO unit/room level addressing where we were able to. This meant that even if one person submitted their response, if another unit that had received a separate code had not responded, our reminder letter and field visits would continue until they did respond. We were able to provide unit-level addressing for around 4 per cent of HMOs.

32. Where we were not able to do this, we assessed the level of within-household response by comparing the response levels in households that contained at least three adults, with at least three distinct surnames, against Electoral Register data on the same basis (3+ adults, 3+ surnames). This demonstrated that the levels of non-response appeared to be no different from the non-response levels in other categories.

33. We also received local-level HMO registrations data for some areas as part of our Local Authority feedback process, which gave specific addresses and the capacity for which they were licensed. We were able to link these addresses to our collected data to demonstrate again that the levels of response (both household and within-household) were in line with other categories of household.

V. Lessons learned

34. Key lessons from the end-to-end use of administrative data in Census 2021 are:

(a) Preparation: it is important to have useful data sources engineered and linked, ready to give you the flexibility to respond to any challenges;

(b) Don’t underestimate the time it takes for preparation of requirements and acquisition of sources;

(c) Data engineering including georeferencing, and matching and linkage are time-consuming and complex;

(d) Prioritizing which bits of data to use from the wealth available can be difficult.

35. We had assessed the definitions and coverage of administrative sources against census definitions in normal circumstances, but then had to re-assess the impact of the pandemic: whether the source was impacted from not being updated regularly, and whether we could

still get access given lockdown working situations, but also whether it still reflected people's actual situations.

36. The importance of accurate linkage: for instance in our AHE model, it was far preferable that we didn't use a source, than that we link the wrong information and make the wrong decision about addresses.

37. These are the main caveats about administrative data source quality:

(a) Currency − not all groups interact swiftly, making records out-of-date and causing under- and over-coverage across geographic areas;

(b) Coherence − be aware of how well sources overlap with what you are measuring in the census, especially if this changes over time (for example due to the pandemic, or changes in the administrative source main use), or is different across areas;

(c) Addressing − variability in addressing practices between sources, and the complexity of certain types of accommodation (notably flats and conversions) makes georeferencing of each address very difficult:

> (i) For instance, some sources would be supplied already georeferenced, and may treat a block of flats under a 'parent' Unique Property Reference Number (UPRN), while another source would use the 'child' UPRNs that nest within the parent one;
>
> (ii) Other sources came in with raw address details that we had to georeference. In some cases, these may use different ways of referring to a block of addresses (eg: flat 1, Block A, Smith Street might be the same address as First Floor Flat, 93 Smith Street);

(d) Our georeferencing tools have improved considerably in recent years, but we are still developing automated processes to understand and clean such complex addressing issues, and have learned a substantial amount from the work that has gone into the census collection and cleaning activities.

VI. Conclusion

38. The use of administrative and paradata sources in Census 2021 was largely as we had planned: we used sources in the preparation of the census address frame and in making non-response follow-up more efficient than in previous censuses. We used sources as we expected to in validating the final estimates of people, households and their characteristics.

39. However, we were also able to make even greater use of sources in ways that went beyond our planned design. We had set ourselves up to react to a range of likely scenarios: we put in place data, methods, resources and governance (see UK Statistics Authority, 2021), and we rehearsed how this might work with a series of simulated exercises. In reality, the situations we faced were slightly different to what we had planned for, but the approach was still appropriate.

40. Response to Census 2021 was the highest we have achieved in recent censuses, and was not impacted as much by the COVID-19 pandemic as we may have feared before collection started. However, we knew that not everyone would respond and that patterns of non-response would need to be properly assessed for any biases, and corrected. Administrative sources have helped us to do this.

41. For information, England and Wales Census 2021 first results were published 28 June 2022 (see ONS, 2022).

References

UNECE, 2018. Approach to using alternative data sources to support the 2021 Census in England and Wales. Paper 15 in UNECE Group of Experts on Population and Housing Censuses, 2018: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2018/Meeting-Geneva-Sept/ECE_CES_GE.41_2018_15-1811158E.pdf.

UNECE, 2021. Progress of the England and Wales 2021 Census. UNECE Group of Experts on Population and Housing Censuses, online, May 2021: https://unece.org/sites/default/files/2021-05/1%20Ghee%20UK%20ENG_2.pdf.

ONS, 2020. Design for Census 2021: https://www.ons.gov.uk/census/planningforcensus2021/censusdesign/designforcensus2021.

ONS, 2021. Students: Census 2021: https://www.ons.gov.uk/census/planningforcensus2021/censusdesign/studentscensus2021.

ONS, 2011. Trout, Catfish and Roach: [ARCHIVED CONTENT] UK Government Web Archive − The National Archives.

UK Statistics Authority, 2021. Methods for 2021 Census Playbook, EAP158: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/11/EAP158-Methods-for-Census-Playbook.pdf.

ONS, 2022. First results from Census 2021 in England and Wales − Office for National Statistics (https://www.ons.gov.uk/).