Methods for producing 2021 Census Microdata

1. Purpose

This document outlines the proposed sampling methods for creating 2021 Census microdata products for England and Wales. The proposed methods differ very slightly from those used for 2011 microdata; the advantages resulting from these changes are outlined in section 5.

The proposed sampling methods result from collaboration between the following teams:

- Census Outputs and Dissemination
- Sampling and Estimation experts within ONS (Gareth Powell and Salah Merad)
- Statistical Disclosure Control (SDC)

The proposed design has been shared with stakeholders who are members of the Census Microdata Working Group. A paper detailing the proposed sampling methods was circulated, and discussions took place in working group meetings in July and November 2019. Working group members are accepting of the proposed method and no concerns have been raised by the group. The working group includes:

- ONS topic experts on migration, travel to work, demography and census transformation, population estimates and statistical disclosure control.
- National Records of Scotland (NRS)
- Northern Ireland Statistics and Research Agency (NISRA)
- Welsh Government (WG)
- UK Data Service
- local authorities
- academia
- market research organisations
- commercial research organisation

2. Action:

The Panel are asked to consider and approve the sampling methods outlined in this paper.

3. Background

Census microdata are anonymised individual level records sampled from a single census; they contain a wide range of individual and household characteristics. Census microdata do not contain any more detail on characteristics than can be gained from standard or commissioned census outputs. Instead, they enable users to study combinations of characteristics and their interactions and perform statistical modelling not possible from the standard outputs. Aggregate estimates are obtained from standard census outputs or commissioned outputs and are produced from the full census database rather than samples of microdata.

Census microdata are used by government, academics, local authorities, research institutes, market research organisations, independent public interest groups and commercial researchers for a wide range of purposes. It is important for analysts using the census microdata to know the sample design.

Some examples of published research using the 2011 Census microdata are:

- P. Troncoso and J.Wathan (2017) Guide to mapping 2011 Census Microdata using R. UK Data Service, University of Manchester. This shows how users can create their own bespoke variables and analyse these under the assumption that census microdata sample are representative of the entire census database.
- D.Wijedasa (2018) The prevalence and characteristics of children growing up with relatives in the UK. Economic and Social Research Council, University of Bristol. Microdata from the 2011 Census were analysed to provide nationally representative, reliable statistics and maps on the distribution and characteristics of kinship care households in the four countries of the UK.
- S.Wilding, D.Martin and G.Moon (2016) The impact of limiting long term illness on internal migration in England and Wales: New evidence from census microdata; This project provides an example of where census microdata have been used for multi-level modelling.

To make data available as widely as possible, and to maximise benefits from the census, we plan to release several different microdata products. These products strike a balance between detail and security and will ensure microdata are available for inquiring citizens through to expert analysts. Consequently, one microdata file will be made publicly available while others will be made available in safeguarded and secure settings.

Census microdata contain no information that could identify a household or individual; direct identifiers (name, address and exact date of birth) are removed, even those in the secure setting. In addition, these products will be created after record-swapping has been applied. The microdata will also include records which have been edited to prevent inconsistent data as well as imputed persons, households, and data values. To protect confidentiality, imputation flags are not currently planned to be included in any microdata file, including secure files; this was also the case for 2011 Census microdata.

3. Census 2021 microdata products

The 2021 Census Outputs team has committed to producing microdata products comparable to those published following the 2011 Census. In addition, following user feedback we plan to produce a safeguarded household file for 2021 and a microdata file that will contribute to the University of Minnesota's Integrated Public-Use Microdata Series (IPUMS) project; an international project which brings census microdata together from over 100 countries. Table 1 details these products.

Table 1: Proposed Census 2021 microdata products

Product	File size	Statistical unit	Lowest geography	Number of variables
Public-access file	Up to 1% of individuals	Persons	Region	Fewer than 20, high aggregation
Safeguarded individual region file	Up to 5% of individuals	Persons	Region	Around 120 (high detail)
Safeguarded individual grouped LA file	Up to 5% of individuals	Persons	Grouped local authority	Around 120 (low detail)
Safeguarded household file	Up to 1% of households	Households	Region	*To be determined*
IPUMS file	Up to 1% of households	Households	Region	*To be determined*
Secure individual file	Up to 10% of individuals	Persons	Local authority	As many as feasible (~200)
Secure household file	Up to 10% of households	Households	Local authority	As many as feasible (~200)

4. Proposed method for sampling 2021 Census microdata and how this differs from 2011

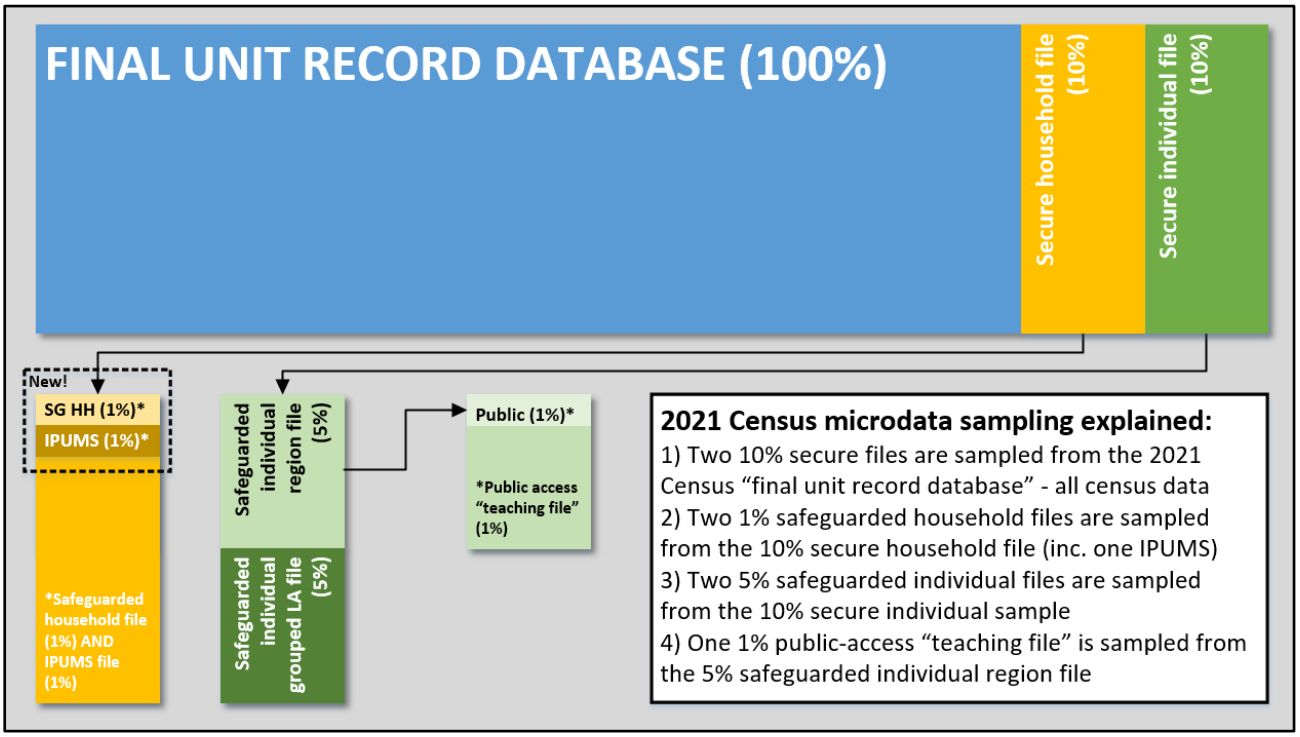
Following discussions with sampling experts within ONS, a slight change to the sampling methods used for 2011 Census microdata is proposed; this change improves upon the methods used for 2011.

2021 Census microdata products will follow the same design principles used for 2011:

- **Nested samples and variables:** Records and variables contained in one sample should 'nest' into larger samples; smaller and less disclosive files should be sampled from larger files.
- **Non-overlapping samples:** Household and individual samples, and samples within each security level, should not contain the same individuals.
- **Systematic random sampling:** The source dataset for samples should first be sorted by local authorities (LAs), then OAs within LAs, then randomised within the OA to avoid clustering. Following this, records are then systematically sampled.
- **Sensitive cases protected:** 'High-risk' records, such as unique individual records and large households, are protected by removal prior to sample creation and designing output categories appropriately; some protection is also received from record swapping which takes place prior to sampling. The exact detail of which records are deemed high-risk is dependent upon the setting in which the data are being released and requires further discussion and agreement with the SDC team. For example, households containing more than 8 individuals are considered high-risk for safeguarded data, consequently they will be removed from the sample frame before the safeguarded household file is sampled.

Given samples should 'nest' within larger samples, the secure household file will be sampled first from the census database, followed by the secure individual file; safeguarded files will then be sampled from the secure microdata files (Figure 1). Prior to selecting the safeguarded samples, high-risk records which need to be excluded from these products to protect confidentiality should be identified and omitted from the sample pool (for example, households of size greater than 8 need to be excluded from the safeguarded household file). The public-access file will be sampled from the safeguarded individual file.

Figure 1: Explanatory diagram detailing how 2021 Census microdata samples will be nested



The proposed approach for sampling the secure household and individual files is outlined below:

2021 Secure household sample	2021 Secure individual sample
1. give random key to all households 2. sort by local authority (LA) 3. sort by output area (OA) within LA 4. sort by random key within output area 5. select **only 1** unique start position; this must be within the first **10 records** 6. from the selected start point, draw every **10th household** 7. continue until the end of the file is reached – the sample will contain around 10% of households	1. give random key to all individuals, filtering out individuals in household sample 2. sort by LA, then by OA within LA 3. sort by random key within OA 4. select **only 1** unique start position; this must be within the first **10 records** 5. from the selected start point, draw every **9th individual** 6. continue until the end of the file is reached – the sample will contain around 10% of individuals.

The secure individual file is sampled after the secure household file is sampled. The safeguarded and public microdata will use the same sampling approach, the only difference being the interval between selected records which will be modified based on the size of the sample.

The secure household file contains approximately 10% of households, this means that around 90% of the census database remains after the secure household file has been sampled (samples are non-overlapping by design). Consequently, to gain around a 10% sample of individuals, every 9th individual needs to be sampled from the remaining 90% of census records.

This approach differs slightly to that used for 2011 Census microdata where:

- After sorting the secure household file as outlined above, 10 unique random start positions were selected from the first 100 records, from each start point, every 100th household was then selected.

- After sorting the secure individual file as outlined above, 10 unique random start positions were selected from first 100 records, from each start point, every 90th individual was then selected.

As in 2011, the sampling method results in equal probabilities of inclusion for all individuals and households, except for high-risk records which are removed to protect confidentiality. Sample weights for households and individuals in each sample are therefore equal and relate to the sample size; consequently, sample weights will not be provided in the microdata files.

Information regarding the sample design for census microdata will be made available to researchers so it can be taken into account when considering inferences they wish to investigate.

5. Advantages of the proposed method for 2021 compared with 2011

The sampling approach proposed for 2021 Census microdata has several advantages over the 2011 approach, with potential disadvantages only existing under unlikely conditions.

Advantages:

- Less likely to result in clustering, as sampling takes place at regular intervals.
- Less likely to skip small OAs, as the minimum number of households in an OA is 40.
- Easier to code, and therefore easier to draw the sample without potential errors.
- No clustering or skipping should result in more representative microdata samples, which should increase the precision of aggregate estimates and the coefficients of a regression analysis.

This approach could be problematic if there is periodicity in the data, but this is very unlikely, since records will have been randomised within each OA.

This alternative approach for sampling 2021 Census microdata is considered to have a negligible impact on the comparability of 2011 and 2021 Census microdata products.

Methods for producing 2021 Census Microdata:

Taking account of comments received from MARP

- A paper detailing the proposed methods for producing 2021 Census Microdata was reviewed by the Methodological Assurance Review Panel (MARP) in June 2021.
- Following this review a small number of comments were received from panel members on the proposed sampling strategy.
- One comment suggested a small improvement to the sampling approach:
    - further stratify the household sample by household size; this should ensure the household samples are more representative in relation to household size
    - further stratify the individual sample by age and sex; this should ensure the individual samples are more representative in relation to age and sex.
- This suggestion has been considered to provide a small improvement; it will therefore be implemented in our sampling strategy for Census 2021 microdata.