Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2021

Estimation and Adjustment Methodology

May 2020

National Records of Scotland

Contents

1. Introduction

1.1 High-level Summary

Scotland's Census aims to gather information from everyone in Scotland. However, it is possible that some people will be missed from the Census or their information will be collected more than once. Estimation and Adjustment is the process that is used to adjust the census data for people who are missed or counted more than once. This allows the final population estimates for Scotland to be as accurate and reliable as possible. The paper explains the process for carrying out Estimation and Adjustment.

1.2 Purpose of this document

Scotland's Census is a household survey of everyone in Scotland which currently takes place once every ten years. For over 200 years, Scotland has relied on the census to underpin national and local decision making. It provides anonymous census estimates which offer a highly accurate picture of the number of people and their characteristics (such as age, health, where and how we live etc.).

National and local government, the education and academic communities, the third sector, commercial business and others use census information in order to plan and provide their operations efficiently and effectively. The information is particularly important when there is no other reliable source or when the ability to cross-reference or compare characteristics of people or households is required.

The census aims to capture details of the whole population of Scotland. However, it is expected that during the census some people and households will be missed. In addition, some people may get counted in the wrong place or more than once. Estimation and Adjustment aims to create a census database fully adjusted for any under or over enumeration at both household and individual level at all geographies.

This document describes the strategy for estimating the level of undercount in Scotland's Census 2021, and adjusting the final database to reflect this undercount. From this database it will be possible to produce tables for any variable, and at any level of geography, that are consistent with each other and accurately reflect Scotland’s population as at census day. 95% confidence intervals will be calculated and published for the main estimates.

The main part of this document describes the estimation (section 3) and adjustment (section 4) elements of the coverage process. The other elements are described more briefly - those prior to estimation in section 2 and those following adjustment in section 5.

2. Secondary data collection and linkage

2.1 Census Coverage Survey

The Census Coverage Survey (CCS) is a follow-up survey that will take place over a five-week period beginning six weeks after census day. It differs from the census in a number of respects, the main ones being:

- it surveys a sample of small areas, covering approximately 1.74% of households in Scotland
- it is conducted by face-to-face interview rather than as a self-completion exercise
- Only a subset of census questions are asked
- It is not compulsory.

The primary purpose of the CCS is to provide an alternative list of households and residents that can be matched against the census to assess coverage levels. The methodology used to assess coverage levels is called Dual System Estimation (DSE) and requires statistical independence between the census and CCS. To help achieve this, the CCS uses a different design and methodology from the census, and there is a limit to the number of field staff that can work on both.

2.1.1 Returns Mechanism

The primary mechanism for capturing responses will be face-to-face interviews recorded on paper questionnaires.

CCS field force workers will return multiple times to addresses to obtain maximum coverage and response rate. In 2021, CCS will be following this process of visiting households up to 10 times and leaving a self-response paper questionnaire on the last visit. The questionnaire and the guidance has been updated in order to improve the quality of these returns. Telephone data capture will be offered as an alternative return mechanism from half way through CCS fieldwork to balance the assumed higher quality of face-to-face interviews with response rate maximisation.

2.1.2 Property listing phase

To ensure independence from the Census, maps will be produced showing streets and houses within each CCS sampled postcode. Interviewers will be required to establish a list of addresses within the sample area prior to CCS interviews commencing. This gives a separate list of properties in the sample area, independent of the Census address list. The postcode address file and other address listing databases cannot be used for this purpose as they are informed by, and used to inform, the Census.

During this property listing phase interviewers record addresses of all properties and indicate if they are residential, non-residential or communal establishments. They will also test the postcode boundaries by going just outside the marked boundary and knocking on doors to ask the post code.

2.2 Questionnaire Design

The CCS questionnaire asks a number of questions both to enable matching with the Census and to support demographic stratification of estimates for estimation and adjustment. The questionnaire is relatively short and allows the interviewer to administer the survey efficiently.

The CCS questionnaire is still to be finalised, however the questions will be similar to those from Census to enable accurate matching. A high-level overview of the proposed questions is given below:

Household level

- Postcode
- Address
- Any usual resident on census night?
- Type of Accommodation
- Self-contained
- Tenure
- Landlord
- Other accommodation
- Number of usual residents
- Number of visitors

Person Level

- Forename
- Last name
- Date of Birth
- Estimated age (if no DOB collected – only for Interviewer version)
- Sex
- Marital Status
- Relation to Person 1
- Student Status
- Term time indicator
- Country of Birth
- Ethnicity (with no write in boxes)
- Activity Last week
- Usual address 1 year ago indicator (not full detail)

- Other enumeration address

Visitor Level

- Forename
- Last name
- Date of Birth
- Sex
- Usual address

2.3 Census to CCS matching

Dual System Estimation requires, within each estimation stratum, the counts of people and households in the census and the CCS, and the number that were counted in both. To achieve this, a matching exercise between the census and CCS will be carried out. In 2011 matching was carried out within each of the 10 Estimation Areas (EA) into which Scotland was divided for estimation and adjustment purposes. In 2021 we plan to conduct matching across the whole of Scotland. Matches that are found between records in different EAs will be noted and the records flagged as out of scope for estimation. This will prevent overcount due to people moving between EAs appearing twice, missed once in CCS and once in Census

The exercise will use a combination of automatic and clerical matching in order to achieve the highest possible accuracy rate. The output for each EA will be a list of census households and people not matched to a CCS record, a list of CCS households and people not matched to a census record, and a list of paired census and CCS households and people which appeared in both datasets.

2.4 Estimation areas

In 2011 both Census and CCS data was grouped into ten geographical areas for all stages of processing from collection to outputs. These groupings served two purposes: to increase the speed of processing by creating more manageable dataset sizes; and, importantly, to decrease the risk of heterogeneity errors in our estimates by estimating the level of undercount for smaller areas with similar levels of non-response. In 2021 improved processing capabilities make processing speed less of an issue, but there remains a need to conduct DSE on estimation areas with homogenous response rates.

The estimation areas (EA) in 2021 will be made up of council areas grouped together based on similarity of demographics related to expected response rate. The council areas making up an estimation areas will not necessarily be geographically contiguous. Estimates are conducted on EA subgroups made by combining each EA with the 5 Hard to Count strata (HtC)[1]. To increase the homogenous response rates within each EA subgroup, non-response follow up for the Census will be prioritised to areas which deviate furthest from the EA subgroup mean response rate.

The exact method for this is under investigation and will be the subject of a future paper.

3. Estimation of the population

3.1 Overview

The process to estimate the number of households and the number of people within them consists of three stages:

1. the application of Dual System Estimation (DSE) to estimate the level of undercount in CCS areas
2. the extension of this to non-sampled areas using ratio estimation
3. the use of small area modelling to derive Local Authority (LA) totals

[1] The HtC index is a scale of 1 (easiest to count) to 5 (hardest to count) which was created to indicate how difficult it may be to enumerate a particular geographical area based on certain demographic features.

A separate, simpler methodology is used to estimate the population of communal establishments.

Once the initial estimates have been calculated, a number of processes are carried out to ensure that the final estimates are as accurate as possible. These take into account known weaknesses in the methodology.

3.2 Stage 1 – Dual System Estimation

Dual System Estimation (DSE) is firstly used to estimate the population within the CCS areas. At this stage people are stratified by age and sex, and households are stratified by tenure.

The use of DSE requires a number of conditions to be met to ensure the minimisation of error in the estimates. These include:

- Statistical independence between the census and CCS - dependence is likely to cause bias in the final estimate. The census and CCS are operationally independent and used different methodologies (section 2.1) in order to help achieve this.
- A closed population - it is assumed that households do not move in between the census and CCS. Clearly this will not be the case, although the CCS is conducted soon enough after census day that the numbers involved should be small.
- Homogeneity - within a cluster, the chance of a person being in the census or CCS is assumed to be the same across all people within the stratum. This is a reasonable assumption since clusters are small and contain similar types of people (the planning areas on which they are based were designed to be internally homogenous with respect to the population). In addition, the main stratification variables group together people and households who would be expected to have similar response characteristics.

- Perfect matching - the matching strategy is designed to achieve a very high level of accuracy.

Census questionnaires received after the beginning of CCS fieldwork are not used in the estimation process, as the presence of CCS interviewers may have affected census response rates in the areas covered.

After matching between the census and the CCS, a 2×2 table of counts of people or households can be derived. This is shown in Table 1.

Table 1: 2×2 Table of Counts of People (or households)

		CCS		
		Counted	Missed	**Total**
Census	Counted	n_{11}	n_{10}	$\mathbf{n_{1+}}$
	Missed	n_{01}	n_{00}	$\mathbf{n_{0+}}$
	Total	$\mathbf{n_{+1}}$	$\mathbf{n_{+0}}$	$\mathbf{n_{++}}$

This output from the matching process will be used to estimate the undercount for each of the sampled clusters. Given the assumptions, DSE combines those people counted in the census and/or CCS and estimates those people missed by using the following formula to calculate the total population:

$$DSE = n_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}}$$

The formula assumes that the proportion of CCS responders that were also counted in the census is identical to the proportion of CCS non-responders who were in the census (this is the independence assumption). This is equivalent to saying that, assuming independence, the odds of being counted in the CCS among those counted in the census should be equal to the odds of being counted in the CCS among those not counted in the census.

The standard DSE formula has been shown to be unstable when the numbers are small, as they will be in some cases. In addition, it does not work in cases where no individuals were matched between the census and CCS. A small adjustment, known as the Chapman correction, is therefore made to the formula as follows:

$$DSE = n_{++} = \frac{(n_{1+} + 1) \times (n_{+1} + 1)}{(n_{11} + 1)} - 1$$

For the main estimates, people will be stratified by sex and five-year age group within each cluster. Households will be stratified by tenure and (where applicable) type of landlord. Estimates will also be produced for several other characteristics, including ethnic group and activity last week.

The output from stage 1 of the estimation process will be estimates of the true population of households and people for the CCS sampled areas, by a number of characteristics.

3.3 Stage 2 – Estimation area population estimates

The DSEs calculated in stage 1 only produce population estimates for the areas in the CCS sample. In the second stage a statistical model is used to obtain estimates for the non-sampled areas.

Of the 32 LAs in Scotland, only Glasgow has sufficient CCS sample to allow a direct LA-level estimate with an acceptable level of precision. The remaining LAs will be grouped together at the estimation stage into Estimation Areas (EAs), which will be the main level at which estimation takes place.

Within each EA, a simple ratio estimator (which uses a straight line of best fit through the origin) will be used to estimate the relationship in the sample between the census count and the DSE. The estimate will be calculated separately for each age-sex group within each Hard to Count (HtC) stratum. This relationship is then used to estimate the total EA-level population of the age-sex-HtC stratum. The variance of the estimate will be estimated by a standard method called the bootstrap.

The output from this process will be estimates of the population for each EA by sex and five-year age group, together with 95% confidence intervals. A similar methodology will be used to calculate an estimate of the number of households by tenure. All of the subsequent stages described below will be consistent with these estimates.

3.4 Stage 3 – Local Authority Estimation

Although the CCS is designed at LA level, practical considerations mean that, with the exception of Glasgow, no LA contains sufficient CCS sample to enable an accurate direct estimate of the population to be made. Nevertheless, estimates of the population (by age and sex) and household count are needed for each LA. The third stage of the estimation process produces LA-level estimates from the EA-level figures.

A small area estimation technique will be applied to produce LA-level population estimates that have lower variances (i.e. smaller confidence intervals) than those that would be produced by just using the sample specific to each LA. The technique uses information from the whole EA to model coverage within the LAs. It makes the assumption that coverage levels across the LAs are the same within each HtC and age-sex stratum. The resulting population (and household) estimates will then be calibrated to the EA estimates to ensure consistency, and their accuracy can also be calculated to provide confidence intervals around the LA population estimates.

3.5 DSE bias adjustments

The application of the DSE at the cluster level is relatively robust to small violations of the assumptions. However, violation of the assumptions can sometimes result in significantly biased estimates of the population. Experience from 2011 indicates that there is likely to be some residual bias in the DSE due to failure of some of these assumptions.

Bias is most likely to result from failure of the independence and homogeneity assumptions. Although there is no way of distinguishing between these two sources of bias, it is possible to make an overall adjustment accounting for both.

Bias can occur both between and within households. Between-household bias refers to a systematic inaccuracy in the estimated number of households, and by extension the total population (since most people are contained within households).

Within-household bias is a systematic inaccuracy in the estimated number of people in each household. This will produce a biased estimate of the total population, even if the estimated number of households is unbiased.

The methodology to correct for between-household bias involves calculating an aggregate number of households for the CCS areas in each HtC stratum of the EA using information from the fieldwork exercise. This method will be discussed in detail in a separate paper.

3.6 Estimates for other characteristics

In order to control the adjustment system (section 4), estimates of the population other than for the age-sex totals are required. The DSE, ratio and LA estimation methodology outlined above will be used to estimate the population by five year age-sex group, ethnicity, activity last week and tenure. As the total person or household estimate will differ slightly for each variable, each set of estimates will be calibrated so that they are consistent with the age-sex total (for person variables) or the tenure total (for household variables).

A different methodology will be used to estimate the household size distribution. This will use a probability matrix approach to estimate the likelihood that a household enumerated in the census as size X is actually size Y. The probabilities are then used to estimate the true distribution. This will be calibrated to the household and person totals (as it can be used to derive both).

3.7 Estimation for Communal Establishments (CEs)

People in CEs are not included in the core estimation process and will be processed separately. This will be the topic of a separate paper.

3.8 Additional adjustments for estimation

In addition to under coverage, there will also be some level of over count in the Census. This happens when the same person is included twice or more in the Census. While some such cases, particularly within households, can be resolved, for many it is not clear which location is correct. The approach for such cases is to apply a down weighting factor to the population estimates to account for over count. This approach to over count correction will be the focus of a separate paper.

National adjustments can be used to correct a persistent issue with the demographic spread in the population estimates that is a departure from comparator sources or other norms. No national adjustment was made to the Scottish data in 2011 census due to the lack of availability of a robust comparator source. The approach to this in 2021 is still under consideration.

3.9 Methods to use when assumptions fail

In the event of the estimates being considered implausible by either estimation analysts or the quality assurance panel, a number of predefined adjustment and improvement strategies are available. The one likely to be used most often is to collapse strata together (e.g. combining adjacent HtC levels or age-sex groups) where one stratum has a small sample size and/or an implausible adjustment level. It is also possible to drop individual clusters from the sample if they are having an undue influence on the final estimates.

Another option is to re-estimate using different strata. For instance, the EA grouping could be changed or the totals could be calibrated to different characteristics (e.g. ethnic group instead of age and sex).

Ultimately, if quality assurance indicates a major failure of the methodology, external data – where it is available and of a suitable quality - can be used to make adjustments.

4. Adjustment

4.1 Overview

Following the production of the population estimates at all levels, the census database will be adjusted to take account of the undercount (and overcount). Synthetic households will be created to allow for those missed by the census, and synthetic people will be imputed into counted households to allow for within-household undercount.

The database thus created will represent our best estimate of the entire population. This adjusted database will be used to generate all statistical outputs from the census.

The outputs from the estimation process define the number of households (by tenure) and people (by age and sex) to be imputed along with basic information about coverage patterns for some other characteristics. However, it is important that we produce plausible values for all characteristics of those households and people missed by the census. The adjustment process can be summarised in three stages.

4.2 Stage 1 – Modelling characteristics

The first stage of the process is to model the likelihood of households and persons, with their characteristics, being missed from the census. These models use the matched CCS/census data to predict (for example) p, the probability of being counted in the census for a 20-24 year old male who is single, white, and living in a privately rented house in the hardest to count stratum.

The models provide the probabilities of both types of undercount: wholly missed households and persons missed from counted households. The variables that are included in the models are those which are controlled explicitly by the adjustment process, which must be characteristics collected by the CCS.

These predicted probabilities are then converted into coverage 'weights' (by taking the reciprocal of p). These weights will be calibrated precisely to the population estimates at LA level described in section 3.4, as these population estimates are the higher quality benchmark.

4.3 Stage 2 – Creation of missed households and people

The second stage of the process will impute people into counted households, and then impute wholly missed households including the people they contain. The coverage weights will determine the characteristics of the synthetic records.

The weights are allocated to each census person corresponding to the likelihood of persons of that type being missed by the census. The census people are ordered by these weights and cumulative actual and weighted counts calculated. The cumulative counts are compared and, if the weighted count exceeds the unweighted count by more than 0.5, a synthetic person is created with the characteristics of the current person.

The characteristics will be limited to those used by the models and those which need to be controlled. Thus a number of 'skeleton' records are created that have certain characteristics. The remaining information will be completed by using the census item imputation system CANCEIS (section 5.2). This ensures the final data is consistent, preserving marginal distributions.

The process for imputing person records is exactly equivalent to that for households.

4.4 Stage 3 – Placement of synthetic households and people

In stage 3 synthetic households (and the people within them) are each placed into a location within the LA, and synthetic people are placed into counted households.

Where possible, synthetic households will be placed into addresses where a household is believed to exist but there is not a complete census record. These include addresses with placeholder forms where the enumerator believed the address was occupied, census questionnaires returned blank, or census questionnaires where only the household questions were completed. The households to be imputed will be compared against the potential locations and scored based on their similarity (including accommodation type and estimated number of residents) to provide the best placement possible.

Where there is no suitable location, a synthetic household record will be created. It will be allocated into a postcode to give it a geographical reference.

The people to be imputed into counted households will be placed into relevant household types: for instance, a baby missed from a four-person household also containing a mother, father and young child would be imputed into an existing three person household of this type. The relationship information will be modified to ensure consistency.

The final output of the adjustment process is a set of skeleton records to be added to the census database.

5. Coverage processes following adjustment

5.1 Post-adjustment imputation

In order to create a complete and consistent database, all variables must have a plausible value. This is achieved by running the post-adjustment database through the item imputation tool CANCEIS. This treats the skeleton records as if they were real people and households who omitted to answer many of the census questions. It imputes values by copying them from a similar donor record, and applies edit rules to ensure that illegal variable combinations are not created.

5.2 Final steps

Before any results can be made public, a further process called Statistical Disclosure Control (SDC) is applied to ensure that no personal data is disclosed in the published figures. This process is outside the scope of this document.

The final output is an individual-level database that represents the best estimate of what would have been collected had the 2021 Census not been subject to undercount (and possibly overcount). Tabulations derived from this database will automatically include compensation for such enumeration issues for all variables and all levels of geography, and will be consistent with the census estimates.

6. Appendix A

List of Acronyms

CCS	Census Coverage Survey
CE	Communal Establishment
DSE	Dual System Estimation
EA	Estimation Area
HtC	Hard to Count Index
LA	Local Authority
SDC	Statistical Disclosure Control

Geography Definitions

Data Zone	The data zone geography covers the whole of Scotland and nests within local authority boundaries.
Estimation Areas	The estimation areas are made up of council areas grouped together based on similarity of demographics related to expected response rate. The council areas making up an estimation areas will not necessarily be geographically contiguous
Hard to Count Index	The Hard to Count index is a scale of 1 (easiest to count) to 5 (hardest to count) which was created to indicate how difficult it may be to enumerate a particular geographical area based on certain demographic features.
Local Authority	Local Authorities are the 32 council areas within Scotland.
Planning Areas	Planning Areas are geographic areas built from groups of postcodes and averaging between 200-400 residential addresses. They nest within Local Authorities.