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

Scotland's Census 2022

Securing high quality Census Outputs and Population Estimates

September 2022

Contents

1. Introduction

This paper provides an update on how National Records of Scotland (NRS) are bringing together the information from Scotland’s Census Collect 2022, the Census Coverage Survey 2022 and administrative data within its statistical estimation methodology to secure high quality Census outputs and population estimates. It includes information on this statistical methodology and how NRS are adapting this to ensure that the Scotland’s Census 2022 delivers the benefit required by its many users.

1.1 Scotland’s Census 2022

Scotland’s Census is the official estimate of every person and household in the country. Data is also collected on characteristics of Scotland’s people and homes. The data collected are processed carefully to ensure they accurately reflect the makeup of Scotland’s population. The 2022 Census collection phase took place between 28 February and 1 June 2022, with late digital and paper returns continuing to be accepted beyond this. More information on the background to Scotland’s Census is available on our website.

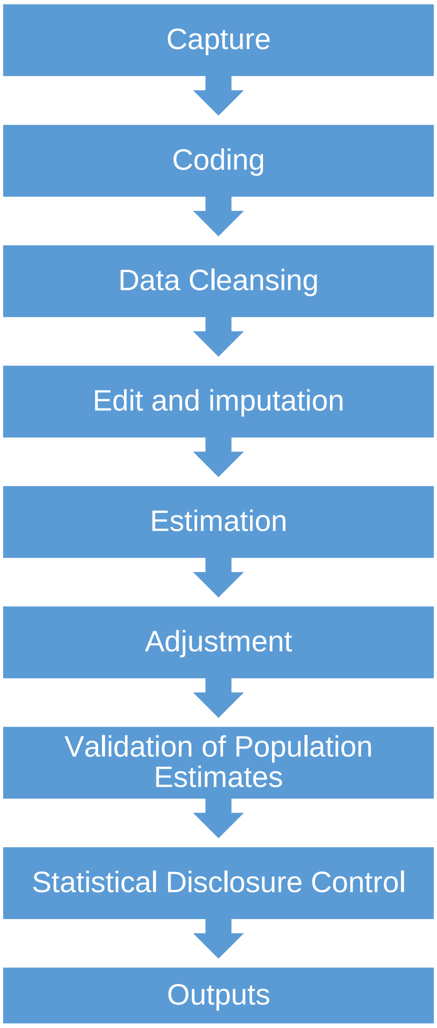
1.2 Census Coverage Survey

A Census Coverage Survey (CCS) is a standard approach used internationally for producing the best possible estimates of the population from a census. The CCS in Scotland was undertaken from 13 June to 22 August 2022. The CCS is the second largest social research exercise in Scotland after the census itself, covering around 1.5% of households. It has been used since 2001 to help estimate how many households and people have not returned a census form and what their characteristics are. By combining this information with administrative data, we can add to the census returns to provide an accurate estimate of Scotland’s population. More information about the CCS is available from the Scotland’s Census website.

1.3 Statistical methodology

Once the responses from census questionnaires are received, they are processed to form a dataset that is, in turn, used to produce high quality statistical outputs. This involves a number of statistical processes that assess, and where necessary, resolve any errors or inconsistencies. This is a large and complex set of tasks that involves numerous statistical methods. Figure 1 shows the basic structure of the processes that form the census data journey.

Figure 1: Census data journey



The purpose of each of these processes is described briefly below:

- Capture – this process captures data provided by census respondents, either online or on paper and turns it into electronic data.
- Coding - assigns each response to a census question a particular code that can be processed to produce census outputs.
- Data cleansing – a collection of processes we apply to census data to account for specific errors, and prepare the data so it’s suitable for later statistical processes. For example, we resolve multiple responses from the same person or household by merging them into one record. This can happen in cases where a person or household submits both an online and paper return.
- Edit and Imputation - ensures the data are complete and consistent at record level. For example, a person may have skipped the occupation question. In this case, we would look at their answers for the industry, number of hours worked and qualifications questions. We can then find a similar response to

these questions on another person's census questionnaire. We can then use their answer for the occupation question to fill in the blank.

- Estimation - produces overall population and household estimates, using information from census returns, the Census Coverage Survey and data from third parties, referred to as “administrative data”.
- Adjustment - creates new records for the missed population and provides a Census dataset for whole population.
- Validation of Population Estimates - We use other data sources that are independent of the census and consult expert panels to ensure our population estimates are correct, accurate and what we would expect.
- Statistical Disclosure Control - prevents the release of disclosive or confidential information about an individual or household.
- Outputs – the process for producing the statistical outputs, supporting information and analysis for release to the public.

More information on each of the processes comprising the census data journey is available on the Scotland’s Census website.

2. Summary of results from the Census and CCS collections

2.1 Census

Scotland’s Census 2022 achieved a national household return rate of 89.2%. In 2011 a 94% person response rate was achieved with the aim for 2022 being to match what was achieved in 2011. The final person response rate will be calculated during statistical processing; it may differ slightly from the household return rate, but will be lower than the 2011 person response rate target.

Another aim for the census was to achieve a person response rate of 85% or more in each Council Area. This aim was met or exceeded in 30 out of 32 Council areas, with 19 out of 32 Council Areas census return rates were higher than 90%. Two council areas were marginally under the 85% aim at 83.3% and 84.0% respectively.

More facts and figures on the census collection are available on the Scotland’s Census website.

2.2 Census Coverage Survey (CCS)

As set out in sections 1 and 2.1 the role of the CCS is to help understand more about the areas and households that did not respond to the Census. The CCS is voluntary and is deliberately weighted to sample a higher proportion of households in the harder to count areas. The CCS completed a response rate of 57.8% which was broadly similar to, albeit slightly lower than, that achieved by the Office for National Statistics (ONS) and reflects the general response rate achieved in the main social surveys carried out in Scotland. Of particular note is that almost a quarter of households included in the CCS sample told interviewers that they would not be responding and should not be contacted again.

3. Statistical methods

Prior to the Census, NRS developed an assured statistical methodology which was designed to take the Census returns from around 94% of Scotland’s population, and 85% in each Council area, and deliver population estimates representing 100% of Scotland. We are now building on this original approach by extending the use of administrative data and evolving our statistical estimation methodology so that we deliver high quality estimates from a lower Census Collect return rate base.

3.1 International Steering Group

The Registrar General for Scotland convened an International Steering Group comprised of census and administrative data experts. The group is comprised of pre-eminent international authorities in census coverage and use of administrative data, and is providing NRS with external advice as to the quality of the census and the statistical approaches that will be used to produce high quality census outputs. More information on the International Steering Group is available on Scotland’s Census website.

3.2 Administrative data

NRS is working closely with partner organisations to secure a range of administrative data. These datasets work in concert for use in the estimation and adjustment processes and NRS current focus includes:

- Flag of last interaction with the health service (Health Activity) from Public Health Scotland,
- Student data from Higher Education Statistics Agency (HESA)
- People registered to vote from the electoral register
- Information on births, deaths and marriages (vital events) from National Records of Scotland
- Data from the School Pupil Census (SPC) and
- People who have been registered with a GP in Scotland, or whose birth was registered in Scotland from the NHS Central Register (NHSCR)

The purpose of these datasets is to provide high quality information on the presence of households and individuals who are missing from the census and CCS datasets, for incorporation into the estimation and adjustment methodology. We will not be using administrative data to identify individuals and names will only be used for linkage purposes.

To produce the data required for use in the estimation and adjustment processes, NRS are exploring linking the Health Activity, HESA and births data to form a “spine”, a list of people and associated information who appear across the datasets. Other vital events (deaths and marriages), electoral register (ER), NHSCR and SPC will inform this. For example, if someone does not appear on NHSCR we might choose to remove them from the spine because they may have left Scotland.

This would result in a “trimmed spine”, which includes records from Health Activity, HESA and births only, with people being removed based on information provided by the other datasets. NRS would then link it to the Census and CCS to work out whether the people in the trimmed spine responded to the Census and/or CCS. This linking will allow us to use administrative data to support estimation, bias correction and adjustment.

3.3 Estimation

After the Census and Census Coverage Survey (CCS) returns have been coded and run through the data cleansing and edit and imputation processes, we will have a clean dataset for both the census and the CCS. We will then run the estimation process, which produces overall population and household estimates, using information from census, the CCS and administrative data.

3.3.1 Previous estimation approach

Our previous approach for the estimation process for census uses the number of records on the census, the CCS and the overlap between them to estimate the total population. For example, if 90% of those responding to the CCS also responded to the census, then we would estimate that 10% of the population did not respond to the census. However, the CCS only covers a small proportion of areas, and extrapolating from those areas to the rest of Scotland can introduce error, especially when we have had a higher non-response rate than anticipated. Furthermore, as the census and CCS both rely on respondents providing data, it may be that the same individuals are more likely to appear on both, as they are more willing to respond to surveys. This overlap leads to bias in the estimates, known as dependence bias, which we can correct by comparing to other data sources such as address lists of properties in Scotland.

3.3.2 Enhancing the estimation process

NRS are planning to change their population estimation methodology for Scotland's Census 2022 to one based on logistic regression, in line with the approach used by the Office for National Statistics (ONS). The current approach incorporates three stages: dual-system estimation (DSE) in strata, ratio estimation and synthetic local authority (LA) estimation. The proposed method condenses this into a single model that calculates record level non-response weights, which can then be used to aggregate up to regional and demographic population counts.

There are a number of advantages to using this approach stemming from its statistical efficiency and ability to pool statistical power from separate regions. These include: reduction in bias caused by heterogeneous response rates; lower requirement for estimation areas (EAs) to be made up of LAs with consistent response rates; less random error due to statistical efficiency of pooling data and the ability to use operational data such as return rates as a predictor of response probabilities.

We are also planning to enhance the estimation process by including administrative records in calculations, either as part of the census or CCS datasets, in place of people who did not respond. Once population estimates have been created, the administrative data would be removed from the census data set. This would provide two substantial benefits to the estimation calculation. The first is that there is a smaller unknown population to be estimated, and so the uncertainty in the size of that population would have a smaller effect on the total population estimate. Secondly, as the administrative data does not rely on data subjects responding to a survey, adding those to the census would reduce the dependence between the expanded census list and the CCS. This leads to less biased population estimates.

We aim to combine administrative data with census and CCS data to create a logistic regression model that can predict the likelihood of a person responding to the census given a set of information, including age, sex, ethnicity and current employment status. We can then use this information to adjust for those who did not respond to the census to estimate the overall total size of the Scottish population, including people who were missed during collection.

3.3.3 Bias correction

We are also planning to use administrative data during our previously planned bias correcting procedures, which are part of the estimation process. These are:

- Alternative household estimate - we will create a logistic regression model to predict the likelihood of dwellings being occupied based on the presence of administrative data, reports from census enumerators, and their activity during the census. Finally, we will use this to create an alternative household count which can be used to compare to the population estimates, identifying and correcting any issues.
- Validation of estimation model - a set of administrative data records, some of which are also census respondents, can be used to evaluate our estimation model, by giving us a target for our estimation of a sub-population (those appearing on the administrative dataset).
- National Adjustment - for this approach, alternative data sources, including administrative data, will be aggregated and compared with the census aggregates. Corrections can then be applied using simple weighing factors for different demographic groups if necessary. For example, this may be used for students, babies and the ratio of males to females.

3.4 Adjustment

In adjustment, new records for the people and households missed by the census are created. Combined with the estimation process, adjustment gives us a census dataset for the whole population.

3.4.1 Previous adjustment approach

Adjustment adds creates new records by either:

- adding people to existing households or communal establishments.
- creating new households in a ‘space’ we already know about. This can be a known address with an occupied property from which we received no response.
- creating new households in a ‘space’ that is not in our address register. In this case we’ll give them a real postcode so we know where they are.

Adjustment is a complex process. For each census record it calculates how likely it is that a similar person or household would be missed from the census. We use this information to choose existing person records to use as donors. Key characteristics from these records are used to create new person records. This is the unit imputation process. This approach was planned as part of the original census design and is consistent with the methods used in other censuses across the UK.

Once the adjustment process has completed, it gives us a census dataset for the whole population that can be used to produce census outputs and population estimates.

3.4.2 Enhancing the adjustment process

We are considering using administrative data to enhance the adjustment process. This would involve using administrative data to improve imputation processes so that the new records created during the adjustment process match non-responding records more closely. These imputation processes select donor records in the census from which to copy missing characteristics. Where we have a person in the administrative data that did not respond to census, we can use the information about their characteristics (age and sex) to better select a similar donor in the census dataset. We can then use administrative records as markers for where these new records should be placed, taking into account the geographical location, age and sex of these records to place a similar record in the correct location. Using the administrative data in this way will allow us to place these records into the correct addresses and reduce the chance of us placing records into vacant households rather than non-responding households.

3.5 Methods used in other countries

Other countries have used similar methods to those being developed by NRS for producing their Census estimates.

For example, in Northern Ireland, the Northern Ireland Statistics and Research Agency (NISRA) used administrative data during their Census Under-Enumeration (CUE) process for their 2011 and 2021 Censuses. The method uses high-quality administrative data to supply demographic information for households/addresses that field staff indicated did not take part in the census. More information is available from the NISRA website.

For the 2018 New Zealand Census, Statistics New Zealand compiled a file of administrative data that provided a good approximation of the New Zealand population and compared this with the census forms data to establish who was missing from the census data. This was then used to add records for real people based on administrative data to the 2018 Census dataset. More information is available from the Statistics New Zealand website.

4. Quality assurance

4.1 Statistical Quality Assurance Strategy

One of the main objectives of the census is to produce reliable data of high quality. To ensure this high level of quality, a number of quality checks are applied to the census data throughout the processing. The term ‘Assurance of Processes’ describes the quality assurance checks that we will undertake at each stage of census data processing. More information is available in our Assurance of Processes Methodology document and in our Statistical Quality Assurance Strategy.

We will also continue adapt our quality assurance strategy; particularly should any changes to the methodology for the estimation and adjustment processes or the extended use of administrative data mean that further quality checks are required.

4.2 Use of administrative data for quality assurance

We will still use administrative data for quality assurance of the 2022 census. Prior to release of the results of the census to the public, we will compare the Census data to other sources, including administrative data sources, to ensure estimates are plausible and any discrepancies are accounted for. We will ensure that the sources used to compare with census estimates are either not being used during census statistical processing, or are used in a way that does not affect the usefulness of the comparison. More information is available in our Validation of Population Estimates Methodology document.

As part of the originally planned census design, we will also use administrative data for quality assurance during the data cleansing phase of processing. Specifically, it will be used to improve the Remove False Persons (RFP) and Remove Multiple Responses (RMR) processes. In RFP, administrative data will help to identify people who would normally be removed by this process, but appear in the administrative data as genuine individuals. Similarly, for RMR, the administrative data can help identify where two or more individuals with similar information (e.g. name, date of birth), within the same household or postcode, are indeed separate individuals, rather than multiple responses for the same individual. Later in processing, where it seems like a person may appear twice in the census dataset across different postcodes, then administrative data will be used to indicate how likely it is that these records represent distinct individuals. More information on these processes is available in the RFP and RMR methodology documents.

Administrative data will also be used to improve the quality of the edit and imputation process by guiding the imputation of age where it is missing or inconsistent in a census record. More information is available in the edit and imputation methodology document.

4.3 Quality Assurance Panels

As part of our original plans, NRS will also arrange quality assurance panels where we will engage with local authorities and topic experts to assist with quality assurance of the census data. This will provide an additional level of assurance of the accuracy of the data. More information on this is available on Scotland's Census website.

5. Census outputs and population estimates

Our data processing methodology, combining the returns received during the census collection with information from the CCS and administrative data, and robust quality assurance procedures allow NRS to be confident of delivering high quality population estimates and census outputs.

5.1 Statistical Disclosure Control (SDC)

The statistical methodology will create a final census dataset for the population of Scotland that will be used to produce census outputs. Statistical Disclosure Control involves both applying statistical methods to the data and controlling access to data and the level of detail that is available to census data users. Before we publish census outputs, we apply SDC methods to the dataset and through the Flexible Table Builder, which will be used to disseminate census outputs, to ensure that individuals and households cannot be identified. These methods help us make sure we are following the rules and laws that protect the confidentiality of census data. More information on statistical disclosure control is available on Scotland's Census website.

5.2 Census outputs and population estimates

Following the completion of the statistical processing, NRS are committed to continue to publish high quality estimates of Scotland's population. These census outputs will provide data at both local and national level that maximises the value of Scotland's Census.

NRS produce annual official population figures for Scotland using data from a range of sources including the Census, registration data on births and deaths, migration estimates, and a wide range of other administrative data sources. NRS continue to improve these statistics through augmentation of existing and new administrative data sources, ensuring the focus is on delivering population estimates that are accessible and valuable for users. Scotland's Census 2022 will provide a new base population, so mid-year population estimates for 2022 will be published after the initial census results are published. The population estimates for mid-2012 to mid-2021 will then be rebased to bring them in line with the 2022 Census population.

We continue to work closely with partner organisations across the UK, including the Office for National Statistics (ONS) and the Northern Ireland Statistics and Research Agency (NISRA), to ensure that UK population data and analysis is coherent, comparable and understandable for all users across the UK.

In preparation for publishing 2022 census outputs we are developing a Flexible Table Builder which will allow users to create their own census tables, increasing the range of data available to users. We have also developed a new website to publish outputs to ensure that the outputs are usable and accessible for users.

6. Next steps

NRS will publish a consultation on census outputs later in the year, giving users the opportunity to provide their views on our plans for the release of census data. This will include the opportunity to feedback on plans for the data we publish, when we publish it and the tools we will use.

NRS will also publish more detailed information on the census statistical methodology in Spring 2023.

Once the statistical methodology has been finalised, we will update our Statistical Quality Assurance Strategy to include any changes or additions to our quality assurance plans.

7. Annex

7.1 Glossary

Term	**Definition**
Adjustment	A process that generates skeleton records that can be added to actual census returns to yield a complete set of census data that reflects as closely as possible the figures produced during Estimation.
Administrative data	Administrative data refers to information collected primarily for administrative (not statistical or research) purposes. This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service.
Assurance of processes	Quality assurance activities at each step of the Census data journey
Census Coverage Survey (CCS)	The Census Coverage Survey is a voluntary, independent, post-enumeration, representative, sample survey used during coverage adjustment to produce population estimates.
Coding	Coding is the process by which the value of a census variable is assigned a code from the responses given by an individual or household.
Council Area (Local Authority)	There are 32 councils (local authorities) in Scotland, the administrative units of local government.
Communal establishment (CE)	A communal establishment is typically managed residential accommodation where there is full-time or part-time supervision of the accommodation.
Data cleansing	A collection of processes applied to census data to account for specific errors, and prepare the data so it's suitable for later statistical processes.

Dual system estimation (DSE)	A statistical method, sometimes referred to as capture-recapture, that uses data from two independent data sources, in this case the Census and the Census Coverage Survey, to estimate the number of individuals and households missed. This estimated number missed along with the number counted during the census allows an estimate to be made of the size of the total population.
Edit and imputation	The detection and repair of gaps or inconsistencies in census data, to ensure a complete and consistent dataset.
Estimation	Estimation produces overall population and household estimates. We'll correct our estimates to account for people who were counted more than once, or counted in the wrong place.
Late return	Any online census return submitted on or after the day the Census Coverage Survey (CCS) fieldwork starts and before the Data Collections Operational Management System is closed, or any paper census return recorded as received by the Royal Mail more than 1 day after the day the CCS fieldwork starts and before the Data Collections Operational Management System is closed.
Logistic regression	A statistical analysis method to model the likelihood of a binary outcome, for example whether an individual or household responded to the census or not or a dwelling being occupied or not, based on prior observations of a data set.
Model	Statistical modelling is a simplified, mathematically formalized method for approximating reality. For example, we will model the likelihood of whether households responded to the census or CCS.
NRS	National Records of Scotland
NISRA	Northern Ireland's census is run by the Northern Ireland Statistics and Research Agency (NISRA).
ONS	The Office for National Statistics (ONS) is the UK's largest independent producer of official statistics and the recognised national statistical institute of the UK. The census in England and Wales is run by the ONS

Quality assurance	Quality Assurance (QA) is about identifying, anticipating and avoiding the problems that can arise from our data inputs or the methods and processes we use to calculate statistics.
Return rate	Return rates are the number of household questionnaires returned as a proportion of the total active household questionnaires that were in circulation (active refers to all households where the address hadn't been deactivated by the field staff during field operations).
Statistical Quality Assurance	Statistical Quality Assurance is about having agreed systems that check and validate the work that we do so that our end product is robust and received well by the end user.
Statistical Quality Assurance Strategy	The implementation of a structured but pragmatic approach to statistical quality assurance
Validation of Population Estimates (VoPE)	The VoPE process compares census data with existing data sources in order to verify that the census data are expected given the comparator data. This process focuses on geographic areas, population groups and topic areas where there are inconsistencies or need for further analysis.