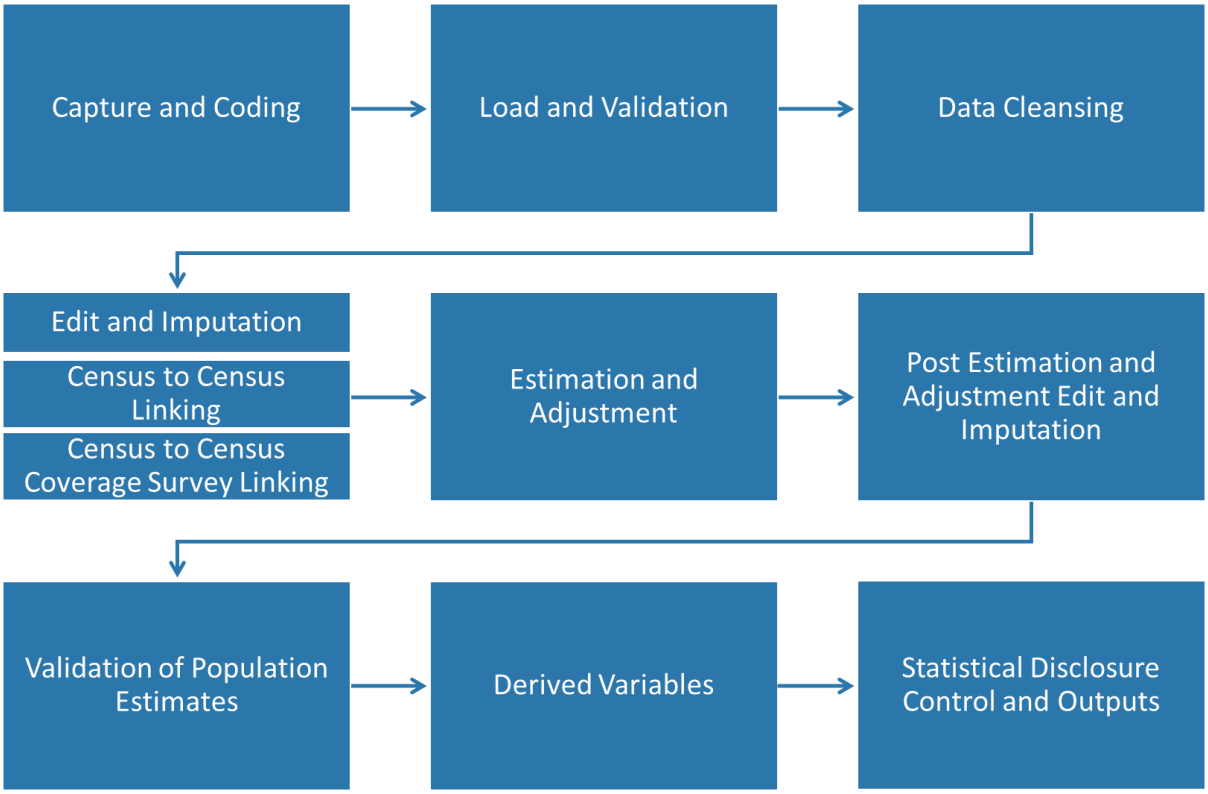
Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2022

Assurance of Processes Methodology paper

April 2021



National Records of Scotland

Contents

1. Plain English Abstract 3
2. Abstract 3
3. Background and introduction 4
4. Methods used by ONS and NISRA 8
5. Proposed method for quality assurance in 2022 9
5.1 Coding 10
5.2 Load and Validation 21
5.3 Data Cleansing – overview 22
5.4 Data Cleansing – Name Reordering 24
5.5 Data Cleansing – Remove False Persons 33
5.6 Data Cleansing – Resolve Multiple Responses 43
5.7 Data Cleansing – Filter Rules 53
5.8 Edit and Imputation 61
5.9 Census to Census Linking 73
5.10 Census Coverage Survey 83
5.11 Estimation and Adjustment 85
5.12 Validation of Population Estimates 94
5.13 Derived Variables 97
5.14 Statistical Disclosure Control and Outputs 103
6. Conclusion 112
7. Annex 113
7.1 Glossary 113

1. Plain English Abstract

Scotland's Census 2022 is the official count of every person and household in Scotland. The census data are used by a variety of stakeholders and data users. 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.

This paper describes how quality checks will be performed on the 2022 Census data.

2. Abstract

National Records of Scotland (NRS)[1] is responsible for planning and carrying out the census in Scotland. Scotland's Census usually takes place every 10 years and collects information from every person and household in Scotland. This allows for the production of a unique set of data not available in any other data sources. These data are widely used by a range of stakeholders and data users for policy development, resource allocation, research and decision-making, at both national and local levels.

The production of high quality data outputs is one of the main objectives of Scotland's Census 2022. Census data processing is built on robust statistical methodologies that include rigorous checks to quality assure the data at each step of the process. Scotland's Census 2022 utilises a number of technological developments and improvements in census processing. This ensures that the quality of the input data and methods of data processing are maximised for the overall quality of the outputs of the final census data.

The census data undergoes a number of statistical processes that form the overall census data journey. This paper details the approach to quality assuring data at each stage of this journey.

[1] National Records of Scotland | Preserving the past, Recording the present, Informing the future (scotlandscensus.gov.uk)

3. Background and introduction

This paper outlines the details of the Assurance of Processes, which is part of the overall Statistical Quality Assurance for Scotland's Census 2022.

The Statistical Quality Assurance strategy[2] for Scotland's Census 2022 provides an overview of methods and procedures for ensuring and assessing that the required high quality of processes is achieved: from the collection and processing of the census data throughout, to the production and dissemination of statistical outputs.

Every household in Scotland must complete and return a census questionnaire and make reasonable steps to ensure that information provided is correct. The digital first approach for Scotland's Census 2022 will encourage the majority of the population to complete their census online. The online questionnaire has built-in functionality that guides respondents through the completion of the questionnaire and helps to minimise the number of inconsistent, invalid or missing responses. This functionality together with the expected increase in online responses will provide higher quality data being submitted compared to 2011 census. The census can also be completed using a paper questionnaire for the whole household or for an individual within a household. Paper questionnaires will be scanned and a specialised optical character recognition technology will convert the written value into electronic data.

There are a number of different questionnaires to account for different types of respondents. The main types of the census questionnaire include:

- household questionnaire – this questionnaire collects data about the household and individuals who live there. This questionnaire is available to complete online or on paper;
- individual questionnaire – anyone aged 16 or over can request an individual questionnaire to complete online or on paper;

[2] For details on the Statistical Quality Assurance strategy, see: Statistical quality assurance | Scotland's Census (scotlandscensus.gov.uk)

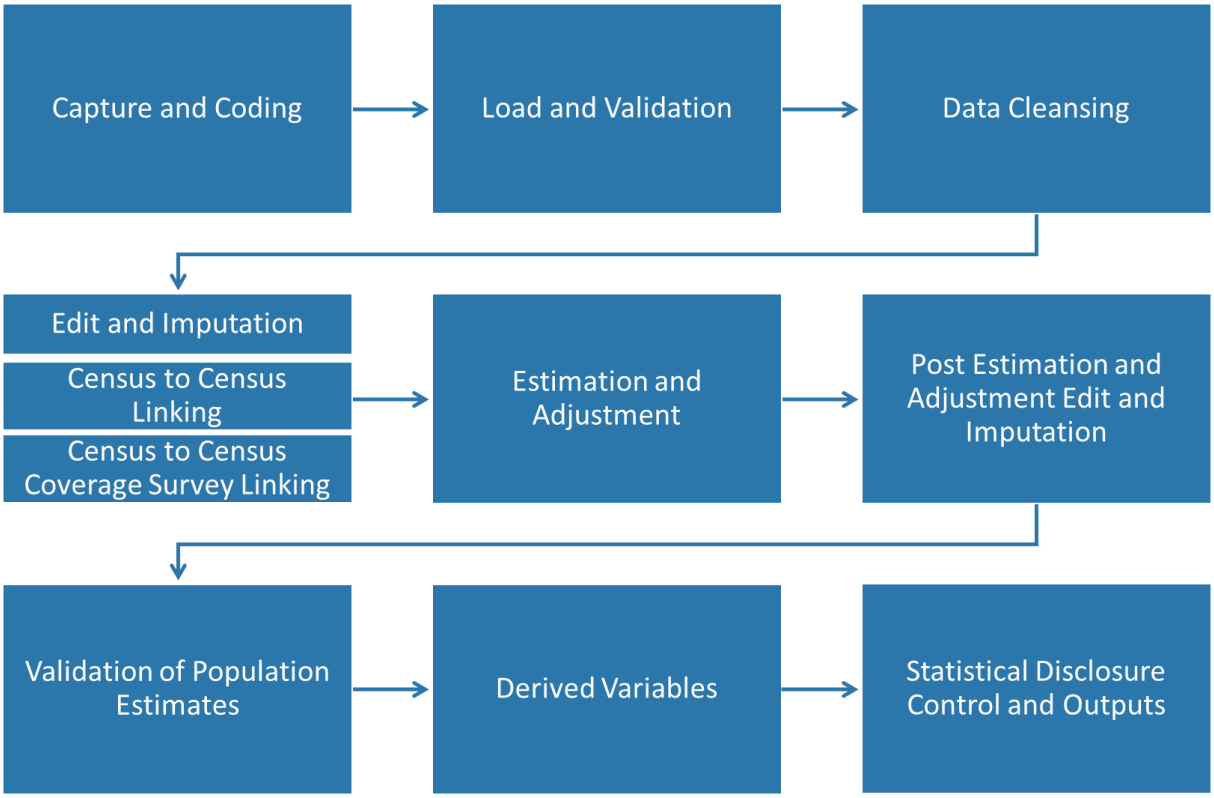
- communal establishment[3] manager questionnaire – the manager of the establishment will complete details about that establishment and the number of people living there;
- communal establishment individual questionnaire – an individual questionnaire to be completed by persons living in communal establishments.

During the census processing all the responses from these census questionnaires are received and processed to form a complete dataset of quality data to produce statistical outputs. This process involves a number of statistical processes that assess and where necessary resolve any errors or inconsistencies. This is a vast and complex process that involves numerous statistical methods. Figure 1 shows the basic structure of the processes that form the census data journey.

Assurance of Processes (AoP) is the name for the series of quality assurance checks that will be performed at each stage of census data processing to ensure that these processes are working correctly and any data issues are captured and resolved. All the quality assurance checks as part of AoP will be performed alongside the census data processing, so that the data will be assured at each stage of the processing before the data can be passed on to the next stage.

[3] A communal establishment is typically a managed residential accommodation where there is full-time or part-time supervision of the accommodation. See Glossary section for more details.

Figure 1. Census data journey



AoP is applied to every statistical process to ensure that during the live census operations, the census data processes are performing according to the developed methodology for each process. The AoP approach involves working closely with statistical teams to understand and assess each process, document the quality assurance checks and establish the systems and reporting standards that will be used to apply these data quality checks. The quality checks aim to reduce or limit the occurrence of errors, and ensure the final data outputs are of the highest quality possible.

The planning of the quality assurance activities that will involve constant monitoring of the statistical processes during the processing of the census data also includes provision for any actions arising from issues identified during the processing. Hence, as part of AoP preparation, for each of the processes described in Figure 1, there is a document that logs potential issues and risks, and lists any actions taken or plans put in place to mitigate these. These documents will be used if any of the quality

assurance checks reveal a potential data quality issue or if any errors in any individual process are detected.

The purpose of quality assurance checks are to identify and resolve any systematic issues that might occur during the running of census data processing. Some quality assurance checks as part of AoP will include manual sample checks that will review individual records. However, these do not intend to check and correct individual census records during this process. This is because the census processing methodologies are designed to identify and resolve any errors and inconsistencies within individual records to create an accurate and complete dataset. The purpose of AoP is to ensure that these processes are performing according to these methodologies.

The Census Day for Scotland's Census 2022 is on Sunday 20 March 2022. This day is used as a reference point to create a snapshot of the population of Scotland. The collection of data will begin earlier to allow everyone to complete their census questionnaire. The online questionnaire will be available for completion from 28 February 2022.

This paper details the quality checks for the census statistical processes. There are a number of similarities between these, however, each process is assigned customised quality assurance checks relevant to the data at each specific stage of processing. These are all detailed in separate chapters of this paper. However, for an overview, the general examples of the AoP checks include:

- summary statistics to access the overall composition of the data;
- trend and data distribution analysis;
- checks with comparative data sources;
- peer reviews of the processes by statistical teams;
- manual sample checks of individual records.

The quality checks as part of AoP are determined in advance, and this paper outlines the high-level methodology of this approach. An automated approach to quality assurance is preferred where possible, as this will reduce the pressure on

resources while affording wider coverage of data, and will reduce any potential bias from manual operation. The proposed automation will utilise the available programming software used for census data processing, including SAS[4], R[5], and Microsoft Excel and Access.

The statistical data processing for Scotland's Census is a vast undertaking due to the size and complexity of the data. Hence, the processing of these data is performed by a number of teams of statisticians, each specialising in a specific process. This ensures that each process is completed by topic experts.

4. Methods used by ONS and NISRA

Office for National Statistics (ONS)[6] and the Northern Ireland Statistics and Research Agency (NISRA)[7] are responsible for planning and carrying out the census for England and Wales, and Northern Ireland respectively.

Each national census office are working closely together to, where possible, develop harmonised methodologies and outputs, which will allow for data comparison across the UK. In cases where the methodology for different processes differs between ONS, NISRA and NRS, these will be highlighted in the relevant section of each chapter.

For more details on the methods used by ONS and NISRA see the respective 2021 Census quality assurance strategies.[8]

[4] SAS (Statistical Analysis System) software is widely used for data management and statistical analysis: SAS: Analytics, Artificial Intelligence and Data Management | SAS UK
[5] R is an open source free software environment for statistical computing and graphics: R: The R Project for Statistical Computing (r-project.org)
[6] For more information see Census - Office for National Statistics (ons.gov.uk)
[7] For more information see Census | Northern Ireland Statistics and Research Agency (nisra.gov.uk)
[8] ONS: Approach and processes for assuring the quality of the 2021 Census data; NISRA: 2021 Census Quality Assurance Strategy

5. Proposed method for quality assurance in 2022

The following section of the paper includes details of every statistical process within the census data journey for Scotland's Census 2022. Each section introduces the relevant process, gives a brief overview, and details the high-level quality assurance approach for each of the processes.

This paper will describe high-level summary for each statistical process needed for understanding of the associated quality assurance that will be performed for each process. Where possible, references to full published papers will be included for detailed methodologies for these processes.

All sections are presented as a standalone description of each process and are structured as follows:

- A brief outline of each process in the style of a short abstract to give an overview of the process.
- **Background and introduction** – introduces each process in more detail including:
    - Method used in 2011
    - Methods proposed by ONS and NISRA
- **Proposed method for quality assurance in 2022** – outlines the high-level methodology of quality assurance procedures for each component of the statistical process.
- **Strengths and limitations** – summarises the main strengths and potential issues of the outlined quality assurance process for each process.
- **Section summary** – a brief summary of the quality assurance approach for each process.

5.1 Coding

For Scotland's Census 2022 every household is required to complete and return a census questionnaire. These responses then need to be captured from the questionnaires and transformed into a digital dataset suitable for further processing. Every answer to each census question is given a numeric code. Possible answers and corresponding numeric values are specified in a specialist document called coding specification. This whole process is called coding.

After the capture process the coding happens either automatically or manually, if the automatic process cannot find relevant value.

The data capture for the online questionnaire is done automatically at the time of completion as the coding logic in embedded in the online questionnaire. In addition, the online questionnaire has a number of built-in functions to aid respondents, and, thus, improves the quality of captured data for coding. However, in some cases manual coding intervention is required if the response is not included in pre-determined coding index.

Paper questionnaires are scanned using the optical character recognition technique to convert hand written information into a digital format. However, different handwritings, or possible imperfections in the scanning equipment may all influence the quality of that capture and the following coding. Similarly, a manual coding process will be applied to these cases.

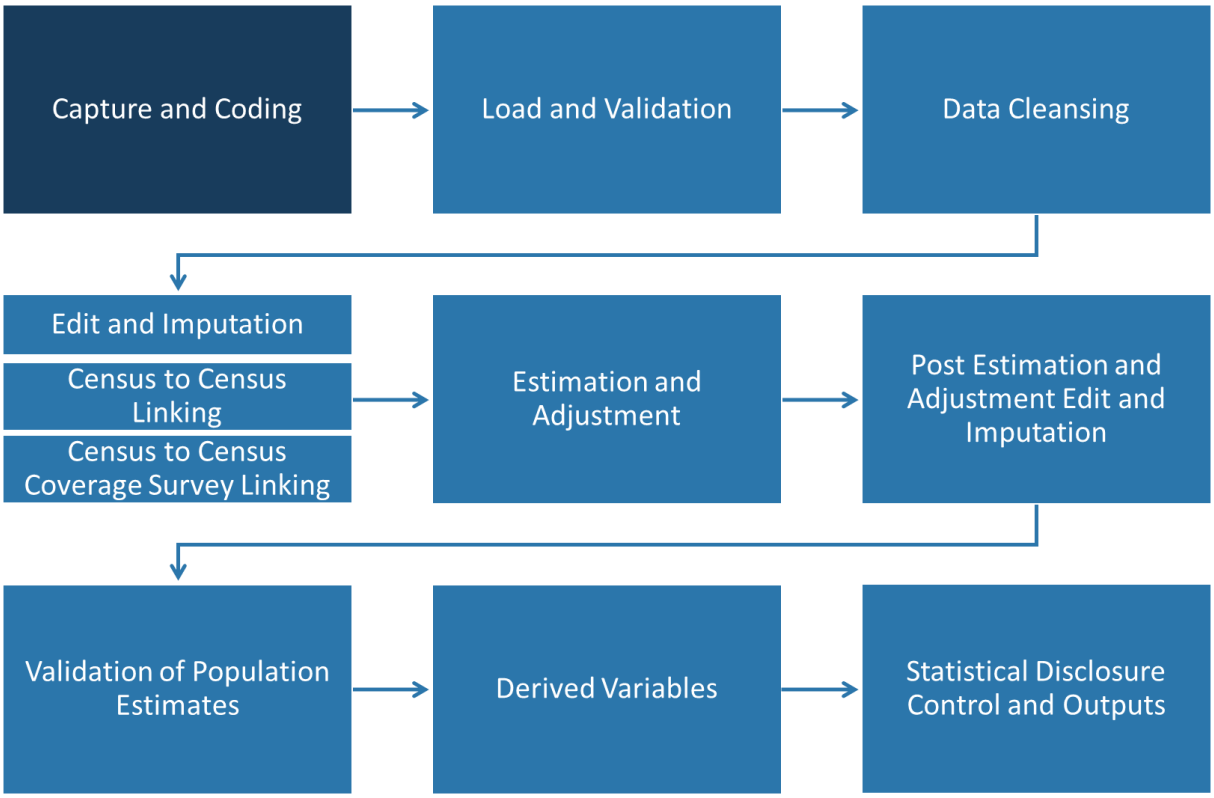
The quality assurance checks for the Coding process include:

- manual sample checks of images of paper questionnaires to check for known scanning and capture issues;
- manual sample checks of the coded responses for paper and online, focusing on the text and address type questions;
- summary statistics of counts and percentages of specific codes, and where possible comparison to existing comparator data;
- secondary independent coding of a sample to assess the accuracy and consistency of the overall coding process.

5.1.1 Background and introduction

After collection of the census responses from online and paper questionnaires, Capture and Coding[9] are the first stages of the census data journey.

Figure 2. Coding process within the census data journey



The Capture process transforms all the census responses into a digital format. This happens automatically for responses completed online. Paper questionnaires are scanned and paper capture technique known as optical character recognition, is used to record the responses into a digital dataset.

The Coding process assigns each response for a census question a particular numerical value or code that can be processed to produce census outputs. Census data processing requires consistently coded data from all types of returns in order to undertake the rest of data processing and provide categorised outputs. For example, the census asks if people are in full-time education. If a person answers 'Yes', their

[9] For detailed methodology, see: Statistical Methods and Data Processing - Coding Solution Methodology Paper | Scotland's Census (scotlandscensus.gov.uk)

answer will be given a code with the number 1. If a person answers 'No', their answer will be given a code with the number 2.

This process uses a pre-determined coding specification, which is an extensive document that specifies possible responses and corresponding numerical codes for each. This process ensures the data is consistent for further statistical processing, analysis and production of outputs. The classification indexes are based on the 2011 census and have been updated during the development work for the 2022 census and following the evaluation of the 2019 Census Rehearsal[10]. These lists include various text responses in relation to the numeric category for a specific response by including different formats and spellings. For example, in the question about Country of Birth a respondent might write in ‘USA or ‘US of A’, both of these responses will be coded to the same numeric code in the classification index under ‘United States’ as they refer to the same country.

The online questionnaire has a number of built-in functions that aid respondents in answering the questions in the most consistent way. This includes automatic skipping of the questions that are not relevant to that person, for example, some questions are asked only of persons aged 16 or over. The online questionnaire also has a type-ahead functionality and drop-down lists that offers respondents lists of suggested answers to some questions, for example, a list of countries for the question about Country of Birth. When a respondent picks an option from the drop-down list, the corresponding code for that option is automatically stored in the dataset. This reduces the burden of coding process, as the system reduces the number of possible incorrect spelling and erroneous answers.

The online questionnaire also includes an address look-up functionality for address questions. This will allow the respondents to search for their address by typing in their postcode or the first line of their address. These features will further improve the quality and consistency of the captured data.

[10] For details of the Census Rehearsal evaluation, see: Census Rehearsal Evaluation Report published | Scotland's Census (scotlandscensus.gov.uk)

The paper questionnaire has guidance within the questions that guides respondents through the questions. However, sometimes respondents might misinterpret the guidance or not follow it and provide inconsistent responses. In addition, the scanning software might be affected by handwriting, which makes it difficult for information to be captured accurately. Hence, there is a higher probability that paper questionnaire responses will require additional manual coding. In addition, without the automatic suggested lists for some questions, responses might be difficult to code. The manual coding process involves specially trained staff who look at the response and make a decision on the appropriate code. A number of additional techniques will be applied to text based questions to correct for possible spelling mistakes and word ordering in order to minimise the requirement for manual coding.[11]

Both paper capture and the online system have undergone thorough testing to ensure quality of capture and coding. In addition, the functionality of the online system has been tested to simulate different types of households to ensure the question routing, validation and the user optimised list cater for the different types of households.

Coding process

The Coding process consists of paper capture, main coding of the online and paper responses, and manual coding.

1. **Paper capture**

   During Scotland's Census 2022, it is expected that a portion of the population will complete their census using paper questionnaires. To be able to process these data, it is necessary for the paper data to be converted into digital data.

[11] For more details, see the paper coding methods section in the Coding methodology paper: Statistical Methods and Data Processing - Coding Solution Methodology Paper | Scotland's Census (scotlandscensus.gov.uk)

After a paper questionnaire is completed and sent back, the data will be captured by scanning the questionnaires. The process utilises optical character recognition technology. This process then automatically codes the data according to the pre-determined coding specification. The original paper scan is retained for archival purposes, and will be used for further quality assurance if required.

2. **Coding – online and paper**

Coding of data is done at the point of input: at the time of submitting for the online questionnaire, and at the point of capture for the paper questionnaire. The coding specifications are the same for online and paper, and the data are merged into one dataset.

At this stage exception codes are assigned when a response is missing, is not automatically matched to the coding index, or is not required to be answered by the respondent (questionnaire routing). Non-exception codes, on the other hand, are answers to the questions, which the system was able to code.

3. **Manual coding**

The data recorded in the text and address questions that receive the exception code (meaning the data item is uncodeable) are sent to manual coding and are manually coded by trained operators.

At the manual coding stage, there are two levels of manual coders: coders and supervisors. If the coder is able to code the item, it is coded and sent back to the main census dataset with a manual coding indicator. If the coder is unable to code the item, it is escalated to the supervisor. Supervisors have more training about the questions, coding techniques and coding indices and will, therefore, be able to code more difficult responses. If the item is coded at this stage, it is included in the main dataset. If the item remains uncodeable it is escalated back to NRS topic experts, who will code the item. If NRS topic experts cannot decide how to code the item, the exception code will be kept

and it will be dealt with further along in the processing at the Edit and Imputation process.

The coding process is applied to different types of responses, including tick boxes, numbers, and text. Most questions have specific coding values assigned to different responses. For example, tick box questions will have a specific codes assigned to each response, or the question about Country of Birth will have a list of current countries with a specific code attached to each.

For questions with a text write-in response option some responses might not match values on the coding index for that question. In these cases, the response will be assigned an exception code and referred to manual coding for further processing. In addition, some questions do not have specific coding specifications. For example, questions on a person's name or the name of the organisation someone works for. Exception codes are still applicable here. Specifically, questions about industry and occupation have the most complex coding indices.

The current approach for the coding process for 2022 Census is to focus on maximising coding at the point of data input and reducing the number of items sent to manual coding where possible. The process will achieve this through comprehensive coding indices and business rules that will reduce exception cases.

Method used in 2011

In the 2011 Census, the same basic principle of data capture and coding was utilised for the Coding process. However, most of the detailed procedures have been replaced and improved for Scotland's Census 2022. This is on the basis of technological improvements and the digital first approach for 2022, where the majority of the population is expected to complete their census online.

These improvements include higher accuracy of data and completeness due to the additional functionality of the online questionnaire, such as question routing,

automatic response validation and built-in coding logic allowing the data to be automatically coded at the time of completion. Further improvements in data capture for paper questionnaires along with the additional coding techniques will also increase the quality of coded paper questionnaire responses. Due to these improvements in both online and paper coding, the requirement for manual coding will be considerably reduced compared to 2011.

Methods used by ONS and NISRA

NRS have worked closely with ONS and NISRA on harmonisation and development of consistent coding lists and indexes. Coding of questions that both NRS and ONS have in common is harmonised across the three census offices. The harmonised coding indices are based on the coding specifications used in 2011. For questions that are either completely different or use different wording or response categories, ONS and NISRA will have different coding specifications.

Unlike ONS and NISRA, NRS will use online drop-down lists for users to pick responses for some questions. This built-in online questionnaire functionality allows for type-ahead functionality and drop-down lists, so that respondents can choose from a list of options. This reduces respondent burden and increases the consistency of coding.

5.1.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the Coding process of the data for Scotland's Census 2022.

1. **Paper Capture sample check**

   To quality assure the paper capture process, a random sample of paper questionnaires images will be compared against the captured and coded data. This sample will include all types of paper questionnaires (including household

questionnaires, individual questionnaires and communal establishment questionnaires) and will focus on any known capture errors, such as mistaking the letter 'I' for the letter 'L'. This quality assurance check will also look for high concentrations of hashes, that is, captured data that was not recognised by the optical character recognition software. If any specific systematic issues are identified, relevant adjustments to the software will be made to improve the process.

2. Manual sample check

A further manual sample check of the automatically coded responses from paper capture and online responses will be carried out. This check will be performed after the manual coding process and will examine all questions. However, it will focus specifically on coding labour market question using SIC and SOC[12], and other text and address based questions, as these types of questions present the most complicated cases for Coding process.

3. Non-exception code check

Summary reports will be produced on the counts and percentages of non-exception codes (answers coded to the coding specification) for each question to assess the overall process. These reports will be produced at each stage of the process (after automatic coding and after manual coding).

These summary statistics will also be used to produce distributions of responses for some of the questions to be compared to data from external data sources. For example, the proportion of census responses coded as 'male' and 'female' will be compared with data from the mid-year population estimates[13]. The census data at

[12] Standard Industrial Classifications (SIC) and Standard Occupational Classification (SOC).
[13] For more details on the data source, see: Mid-Year Population Estimates | National Records of Scotland (nrscotland.gov.uk)

the Coding stage of the processing has not undergone Data Cleansing procedures, hence, it is not expected the distribution comparison to be a very close match to the existing data. This comparison will be used as a tool to identify questions or response categories where there are large differences between the census and comparable external sources.

4. Exception code check

Summary reports will also be produced on the counts and percentages of exception codes (answers that the system was not able to code) for each question. Similarly, these reports will be produced at each stage of the process (after automatic coding and after manual coding) to assess the overall Coding process.

5. Hard to code check

A manual check will be carried out of the responses against specific coded values that are known to occasionally be coded incorrectly. There is a pre-determined list of potential cases, however, additional values may be added during census operations as the live data is received. Any changes and the resulting actions will be recorded and reported for quality assurance.

6. Manual coding report

A sample of the data after manual coding stage of the process will be assessed. The sample will include the data that were initially coded as uncodeable and were sent to manual coding for further processing. This is a sense check of these data after the manual coding process has assigned relevant codes. This check will mostly focus on the text and address type questions.

7. Accuracy check

To assess the accuracy rate of the overall coding process, a sample of data that have been successfully coded will be sent to the manual coding operation to check for accuracy. The sample will include data from both questionnaires types (online and paper) and will cover the breadth of questions asked through the census. The sample data will be coded by manual operators without the knowledge of the original coded value, and then compared to the initial coded values using an automated process. This will demonstrate the rate of accuracy and consistency of the coding process.

8. Manual sample check

A final manual sample check of data will be completed across all questions. The check will group similar topic questions together for ease of application, and will focus on the groups of text and address based questions, taking a bigger sample size of these compared to smaller samples of tick based questions that are generally easier to code. The coded data samples will be compared to the values in the coding specification. The results will report on the overall accuracy of the Coding process.

5.1.3 Strengths and limitations

The main concern about the planned approach to quality assurance of the coding process is that the manual sample checks are resource intensive. This will require careful consideration of available resources and a suitable sample compilation.

However, the digital first approach for Scotland's Census 2022 will encourage the majority of the population to complete their census online. The online questionnaire guides respondents through the completion of the census questionnaire and includes

a number of built-in validation rules and messages to minimise the number of inconsistent, invalid or missing responses. This functionality together with the expected increase in the proportion of online responses will provide higher quality data being submitted into the coding process compared to the 2011 Census, which in turn will increase the data quality for further processing.

5.1.4 Section summary

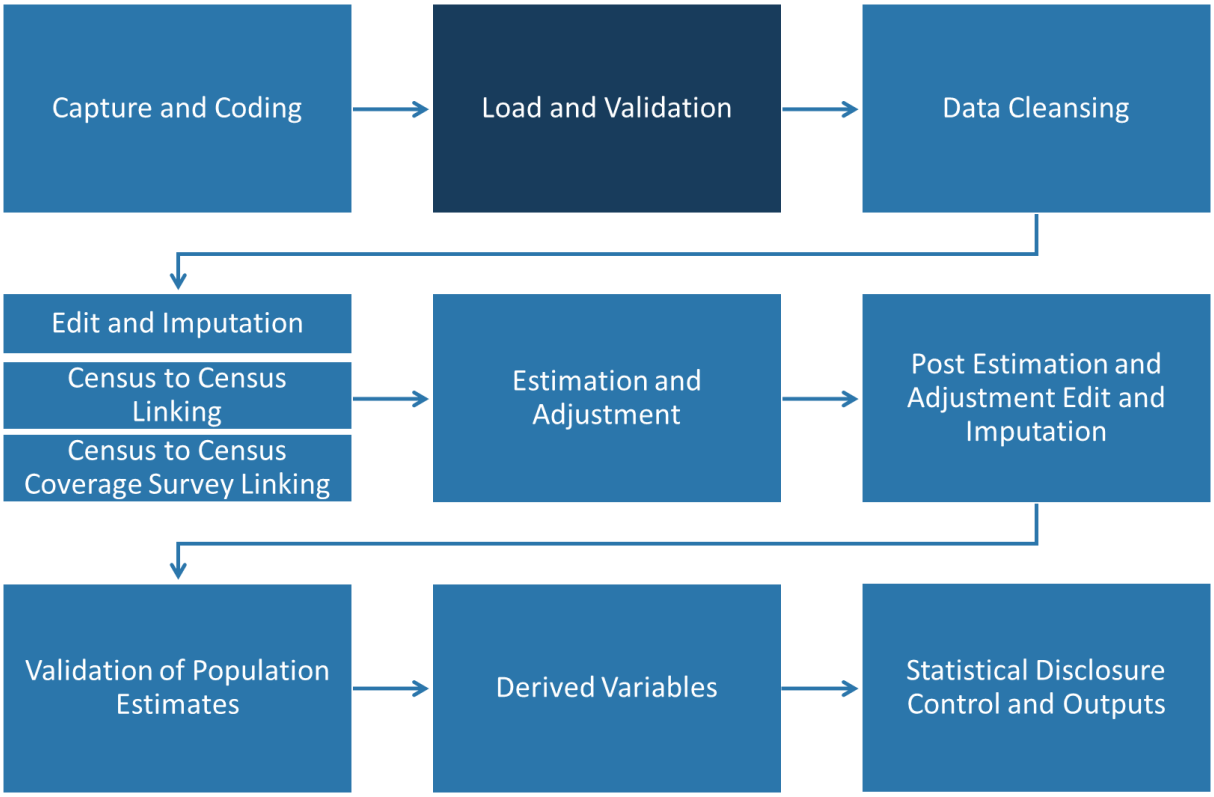
To quality assure the Coding process during Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.

These include a number of manual sample checks of the captured and coded data, distribution checks of the different codes, and an accuracy coding check.

5.2 Load and Validation

Load and Validation is the process of receiving data after the data have been collected and after the Capture and Coding stage. The data are received into the secure environment in a regular feed from the start of collection until six weeks after the Census Day.

Figure 3. Load and Validation process within the census data journey



The process in itself is not statistical in nature. The quality assurance for this process will focus on ensuring the data are received in the correct format and adheres to pre-determined specification and processing rules. The process of Load and Validation has been evaluated following the 2019 Census Rehearsal and is reflected in a Census-wide data management and control plan[14].

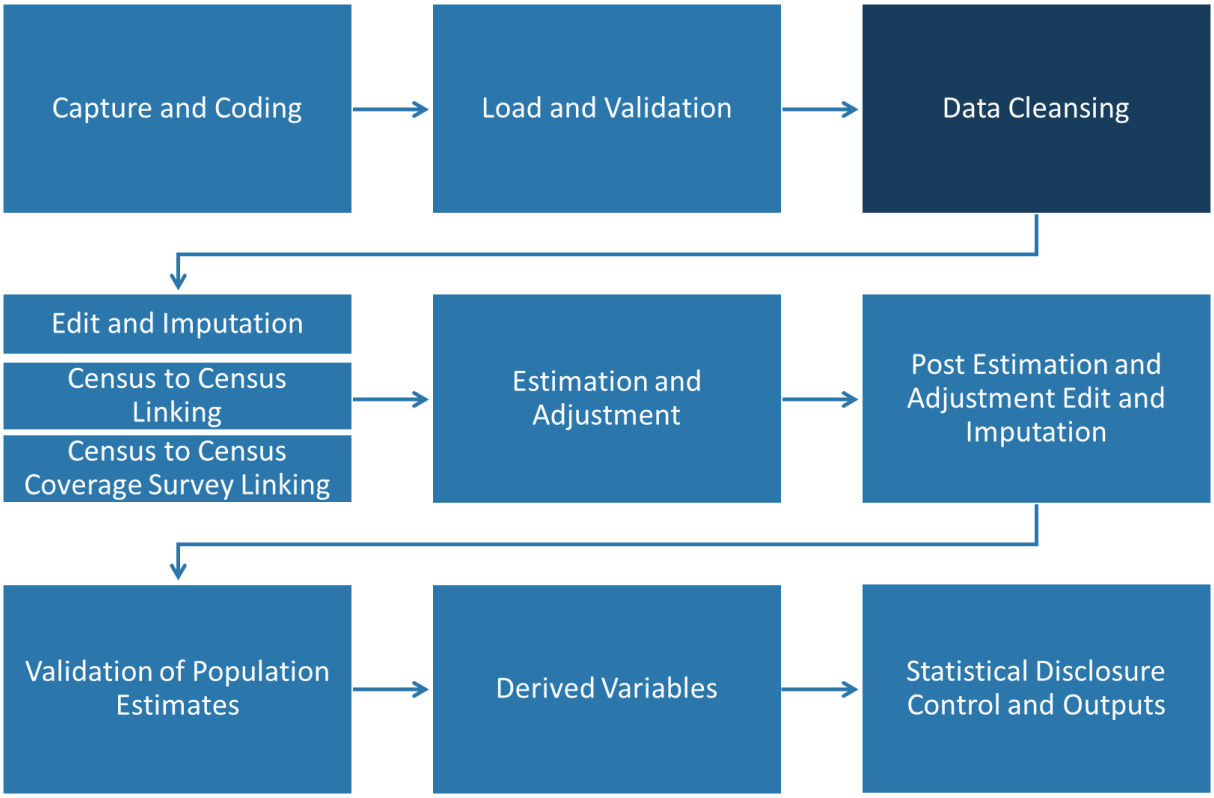
[14] For more information on assurance within the census programme see: Census programme assurance | Scotland's Census (scotlandscensus.gov.uk)

5.3 Data Cleansing – overview

The Data Cleansing stage of the census data journey includes processes that identify and resolve specific errors and inconsistences, and prepare the data so it is suitable for later statistical processes.

Data Cleansing includes a number of separate processes, which are described individually in the following chapters (Sections 5.4, 5.5, 5.6, and 5.7). This paper focuses on presenting the quality assurance approach of each census data processing stage, hence, describes these statistical processes at high-level of detail. For full details, see published methodology papers presented for external assurance as part of peer review: Peer review and governance | Scotland's Census (scotlandscensus.gov.uk).

Figure 4. Data Cleansing process within the census data journey



The Data Cleansing includes the following processes:

- **Name Reordering**[15] (Section 5.4) – this process compares the names from the person's individual name question to the names in the relationship question. The process identifies any discrepancies in the order the names have been entered. The process then suggests and implements the ordering that minimizes these discrepancies.

- **Remove False Persons** (RFP)[16] (Section 5.5) – this process looks at spurious person records in census dataset. The process identifies and resolves any names that might suggest the record does not relate to a genuine person. In addition, the process applies the condition of minimal required variables for the census record to be considered as a valid record for an individual.

- **Resolve Multiple Responses** (RMR)[17] (Section 5.6) – this process identifies and resolves duplicate records of communal establishments, households and people that inevitably appear in the dataset.

- **Filter Rules** (Section 5.7) – this process resolves issues and inconsistent information in the answering path (routing) of a questionnaire.

[15] For full methodology details see: PMP005: Name reordering methodology | Scotland's Census (scotlandscensus.gov.uk)
[16] For full methodology details see: PMP011: Remove false persons - methodology | Scotland's Census (scotlandscensus.gov.uk)
[17] For full methodology details see: PMP014: Resolve multiple responses - identify duplicates | Scotland's Census (scotlandscensus.gov.uk)
PMP021: Resolve Multiple Responses Prioritisation and Resolution: Peer review and governance | Scotland's Census (scotlandscensus.gov.uk)

5.4 Data Cleansing – Name Reordering

The household section of the Scotland's Census 2022 paper questionnaire asks the names of everyone living in a household and how these people relate to one another. Further, the questionnaire asks individual questions about every person in the household. The order in which people appear on the individual part of the questionnaire should match the order used in the household section, so that the relationships within household are recorded correctly. In the cases where the individual sections appear in a different order to the relationship question, a process of name reordering compares the names for each household record and matches the names to correct the order.

Name Reordering is part of the Data Cleansing step of census data processing. The Name Reordering process is applied to household census records that might appear on the census paper questionnaire in a different order to that recorded in the relationship question. The process preserves the correct relationship between the household members by reordering the person number for individual records to align with the order recorded in the relationship question.

The quality assurance of the Name Reordering process includes confirming that the deterministic algorithm delivers the decisions in accordance to the Name Reordering methodology.

For each census record the Name Reordering algorithm assigns one of the following options:

- records do not require reordering;
- records passed on to automatic reordering;
- records passed on to clerical review.

The Name Reordering process was not used in Scotland's Census 2011. This is due to the difference in the design of the paper questionnaire that allows to record the names on the relationship question.

5.4.1 Background and introduction

The functionality of the online census questionnaire for Scotland's Census 2022 allows for the names of the people from a household to be entered once and then to be automatically piped throughout the questionnaire. This functionality removes the need for the householder completing the questionnaire to enter the details of the household members in a specific order, as their names will appear on each relevant page of the online questionnaire.

Scotland's Census 2022 paper household questionnaire contains three questions that record the names of the people living in a household. These include the relationship question where the information on how people in a household relate to one another. It is important that the individual questions for each person in the household match the order they appear in the relationship question, so that the relationships within the household are recorded correctly.

The instructions in the paper questionnaire ask to write the names of household members in the same order throughout the questionnaire. For most cases it is expected that this will be followed and the individual person questions for people in the household will be answered in the same order as the people's names on the relationship question.

However, in some cases the order might not be followed. Therefore, the process of name reordering is applied to compare the names on the individual person questions against the relationship question. In cases where the names do not match, the algorithm attempts to reorder the individual responses where possible to account for the error. This process ensures that the correct information on the relationships between household members is preserved in the census data.

For example, if in the relationship question the name for John Smith (father) was entered as Person 1, and Bobby Smith (son) as Person 2, and later on in the individual section of the questionnaire the details for Bobby Smith were entered first as Person 1, and John Smith was recorded as Person 2. In this case, the record for Person 1 will not match the information in the relationship question, as the 'father'

will appear to be younger than the 'son' is, because the information for these two people was entered in a different order (see Table 1 below).

Table 1. Example of name order mismatching between household relationship question and individual questions

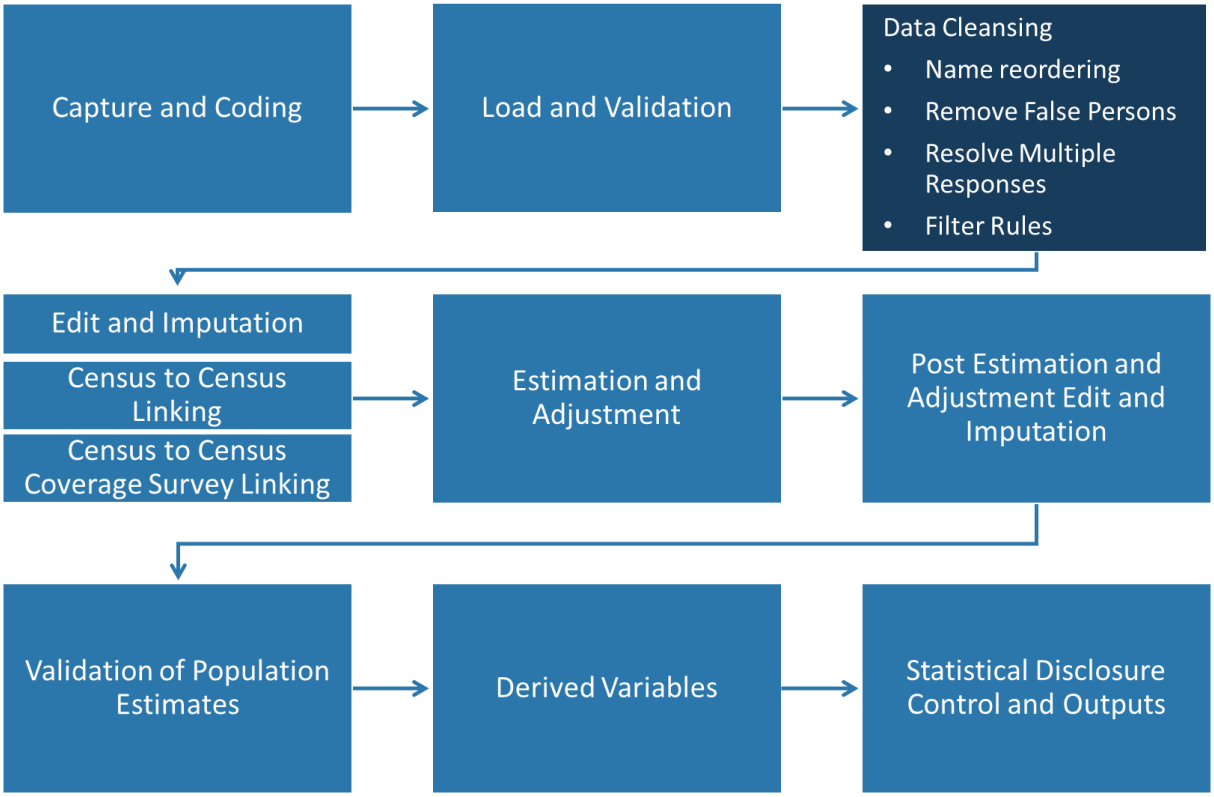
Household questions			Individual questions			
Person number	Name	Relationship question	Person number	Name	Date of birth	Age
1	John Smith	Father	1	Bobby Smith	01/01/2021	1
2	Bobby Smith	Son	2	John Smith	01/03/1986	36

The Name Reordering process will identify this mismatch and will propose the reordering of the two records. Following the Name Reordering process, John Smith (father) will be recorded as Person 1 and Bobby Smith will be recorded as Person 2, both in the same way as in the relationship question.

Table 2. Example of the records after the Name Reordering process

Household questions			Individual questions			
Person number	Name	Relationship question	Person number	Name	Date of birth	Age
1	John Smith	Father	1	John Smith	01/03/1986	36
2	Bobby Smith	Son	2	Bobby Smith	01/01/2021	1

Figure 5. Name Reordering process within the census data journey as part of Data Cleansing



The Name Reordering process[18] compares the names from the person's individual name question to the names in the relationship question. The algorithm then scores the comparison taking into account nicknames, phonetically similar names, names that agree at the start or the end, and also according to a character by character comparison. The scoring is then used to find the ordering that minimizes discrepancies.

The name reordering algorithm can have three outcomes:

1. Algorithm cannot find a better scoring – there is no need for reordering. The majority of cases that have the correct names order will fall into this category.

[18] For full methodology details see: PMP005: Name reordering methodology | Scotland's Census (scotlandscensus.gov.uk)

2. Algorithm finds reordering with a high score – these records will be reordered automatically.

3. Algorithm cannot decide in case of a low score how the records should be reordered – these records will be passed on for a clerical review.

The testing on Census Rehearsal 2019 data suggests that approximately 1,000 out of 2 million household records would require their names to be reordered.

The census records with inconsistent number of individuals in the household, or missing individual records, will result in the algorithm producing a very low name matching score. There will be limited information to suggest a better order and, therefore, these records will be kept unchanged.

However, it is worth noting that the missing person's individual records on paper household questionnaires might be occurring more frequently in 2022 compared to 2011 due to potentially more people requesting individual questionnaires. People requesting an individual questionnaire will be included in the household questions; however, the householder will be instructed to leave the individual questions blank for these persons. This applies to the census paper questionnaires only.

Method used in 2011

The Name Reordering process was not used in 2011, as the names in the relationship question were not captured for processing. This was due to a different design of the paper questionnaire. See Figures 6 and 7 for illustration.

Even though Name Reordering process was not used in 2011, some cases where the persons appeared in different order were resolved manually, when incorrect ordering led to conflicts in the data. This could occur, for example, if a person did not appear to be younger than their parent was. However, this was a time consuming process, and could only detect cases that were obviously incorrect.

The new design for 2022 allows the names to be recorded part of the census dataset. In addition to the Name Reordering methodology, it provides a level of confidence that the relationships between household members are recorded correctly in the census. This information is important for understanding the individual characteristics within the household composition.

Figure 6. Scotland's Census 2011 household relationship question (paper questionnaire)

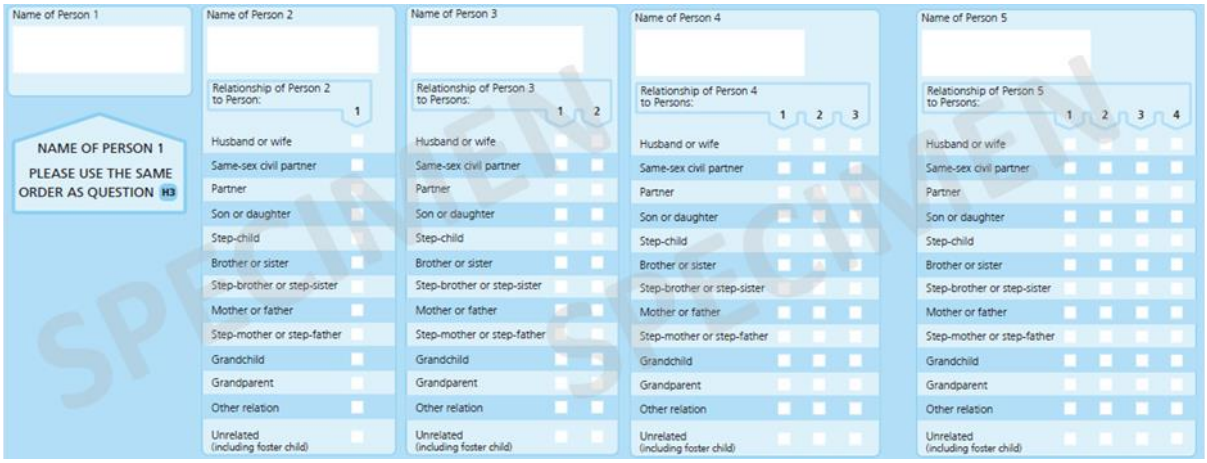
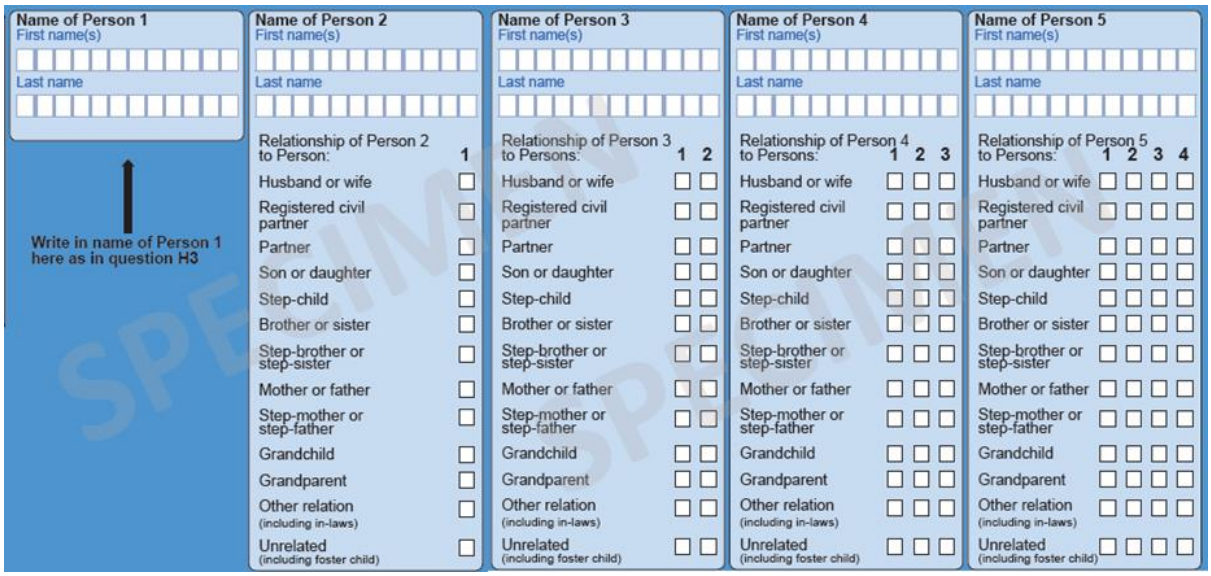


Figure 7. Scotland's Census 2022 household relationship question (paper questionnaire)



Methods proposed by ONS and NISRA

The design of the relationship question for paper questionnaire for ONS and NISRA for Census 2021 is similar to the 2011 version. For this reason, the names in that question are not captured to be used in data processing. Hence, the name reordering process is not used for their census data processing.

5.4.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Sample check**

   For some census records, the process algorithm will find a higher scoring ordering and will select these records for an automatic reordering. A summary statistics report will demonstrate the overall scale of this process.

   In addition, a manual check on a sample of these records will be performed to ensure that the records were reordered correctly by the process. The records will be assessed according to the reordering rules set out in the main process methodology.

   It is anticipated that there will be around 1,000 cases of household records to be assigned to an automatic reordering in the whole census dataset. This quality assurance check requires a manual review and, hence, is resource intensive. Because the process follows a pre-determined algorithm, only a small sample of records will be sufficient for this manual check to reveal any issues or inconsistencies in the process of assigning the records through automatic reordering.

   Further, another manual check will be performed on a sample of census records that were not selected for automatic reordering. These are the records that the algorithm has not been able to find a better scoring reordering.

These quality assurance checks will confirm that the main process algorithm assigns correct scores and a correct relevant action (automatic reordering or clerical review).

2. Clerical review

The name reordering algorithm will identify some census records as having a high enough score to suggest reordering, however, the score will not be high enough for an automatic reordering. In this case, these census records will be passed on for a clerical review.

Clerical review involves an additional evaluation that can have three possible outcomes:

- accept reordering suggested by the algorithm;
- keep the original ordering;
- suggest a different reordering.

To quality assure the clerical review process, a sample of the clerically reviewed records will be reviewed a second time by a different reviewer. If the decisions from the sampled clerical review matches those from the main clerical review, this will demonstrate the validity of the overall clerical review process, and hence, the accuracy of the reordering process for more challenging cases.

3. Summary statistics

Summary statistics reports will be produced to demonstrate the extent of the records that are selected by the process algorithm for clerical review and the outcomes of those reviews together with the original scores. This will reveal any unexpected or excessive numbers of census records undergoing clerical review.

5.4.3 Strengths and limitations

The potential scale of Name Reordering process has been estimated using the results from the 2019 Scotland's Census Rehearsal. The Rehearsal data provided a suitable representation of the results expected in the Census 2022. However, there is always a possibility that the data during the census will reveal different patterns and that the name reordering corrections will need to be run at a larger scale than previously expected. This will be most relevant for the clerical review process since it requires additional resource for manual correction.

However, the digital first approach for Scotland's Census 2022 will encourage the majority the population to complete their census online, which will reduce the number of census responses submitted on paper. The online questionnaire has functionality to aid respondents in completing the questions in the correct order, and automatically includes the name of the household member on individual questions. This functionality together with the expected increase in online responses will reduce the number of paper questionnaires and in turn minimise the need for Name Reordering process.

5.4.4 Section summary

To quality assure the Name Reordering process for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.

These include a manual sample check to ensure the name reordering algorithm assigns correct scores to household records suitable either automatic reordering or clerical review. The clerical review process will be quality assured by completing the clerical review process twice on a sample of records. Matching the results from this exercise will confirm the validity of the overall clerical review. Further, a summary statistics report will be produced following the name reordering process to ensure it has produced expected results.

5.5 Data Cleansing – Remove False Persons

Every household in Scotland must complete a census questionnaire and make reasonable steps to ensure that information provided is correct. However, some of the information provided in census responses might be false. This might be due to a respondent not taking reasonable steps to ensure that the information provided is correct either intentionally or unintentionally. This can also occur if the scanning process picks up accidental marks on the paper questionnaire as answers. Such issues can falsely increase the number of people counted in the census (often referred to as overcount). The remove false persons (RFP) process, where possible, identifies and removes such responses to prevent them from being passed on for further data processing, ensuring that only genuine individual records are processed.

Generally, these issues affect a small portion of census responses, as most of them arise from errors in scanning of paper questionnaires. The majority of responses in Scotland's Census 2022 will be online, hence, it is expected that the functionality and built-in validation in the online questionnaire will significantly reduce these issues compared with the 2011 Census.

The quality assurance checks for the RFP process include:

- a summary statistics report on records information which is identified as false for the following variables: name, date of birth, sex, marital status, household relationship;
- clerical review on a sample of records that are classed as borderline pass cases to ensure that the passed individual records are genuine.

5.5.1 Background and introduction

Every household in Scotland must complete a census questionnaire and provide information that is full and accurate. In some cases the information received does not related to a genuine person response, and the RFP process identifies and accounts for this information in the census data before further processing.

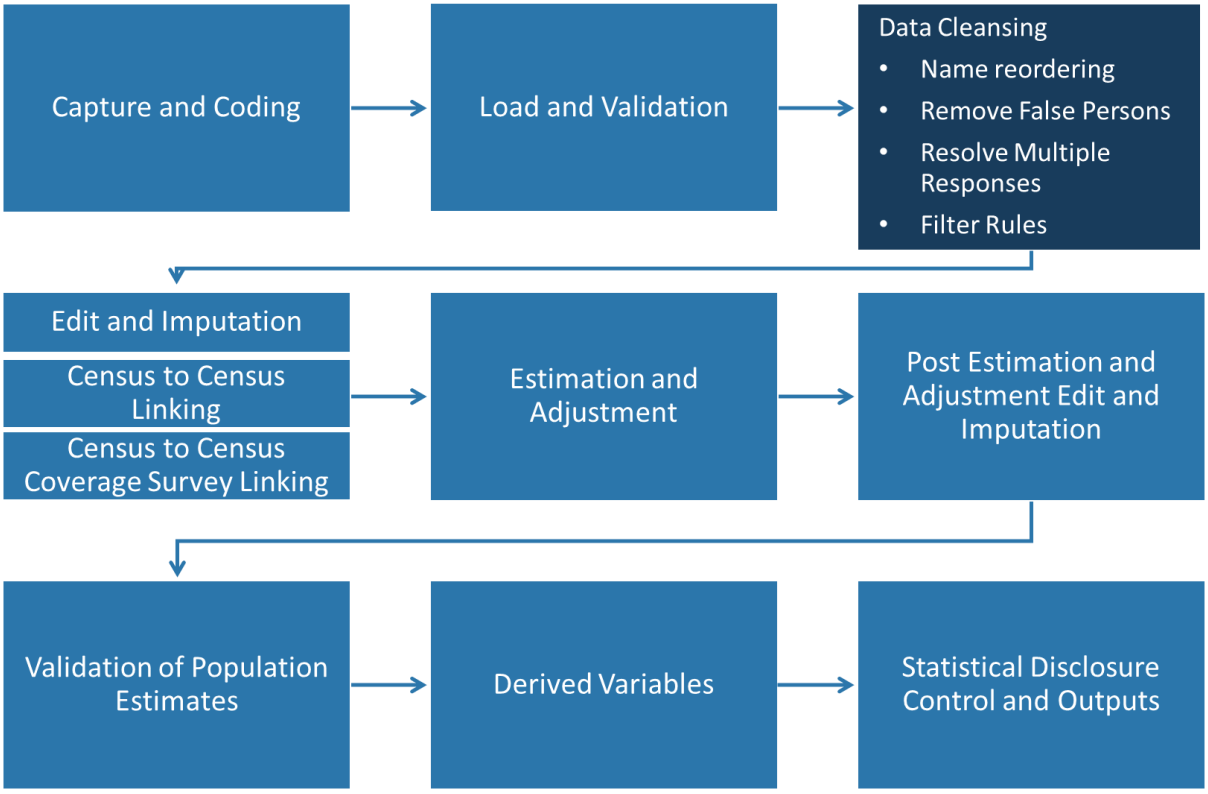
Sometimes the false responses occur when respondents intentionally or unintentionally submit a response that does not relate to a genuine person. For example, someone might write on the paper questionnaire 'No one lives here' in the person questions for an empty property. This will be picked up at the scanning and data capture stage as a response to be passed on to Data Cleansing.

Many of the specific issues such as dust marks during scanning and capture are now resolved through the improvement of the capture technology since the last census. Moreover, fewer responses on paper questionnaires are expected for Scotland's Census 2022 as the completion will be primarily conducted online.

Remove False Persons (RFP) is a data cleansing process that looks for spurious person records in census data. The need for a RFP process still exists as respondents often return paper questionnaires without enough information to determine if it represents a real person.[19]

[19] For detailed methodology, see: PMP011: Remove false persons - methodology | Scotland's Census (scotlandscensus.gov.uk)

Figure 8. Remove False Persons process within the census data journey as part of Data Cleaning



The RFP process consists of a number of interim steps as follows:

1. **Check for false names** – this is a new method for Scotland's Census 2022. This filters for obviously falsified names and other indicators that a census record is not actually a record of an individual person. This step identifies blatantly false name information, for example, 'N/A', 'NO ONE IS HERE'. This also includes the cases where the respondents purposefully obscure their names, such as 'ANON' or 'ANONYMOUS'.

2. **Remove false names** – this step removes records that are confirmed to contain blatantly false name information only, as identified during the check for false names step. For example, if a record contains information only for the name question as 'NO-ONE', and no other information, the record will be removed. However, if a record contains an obviously false name, and other

information, this record will be passed to the next stage. This could be that a respondent purposefully obscured their name only but completed the rest of the census questionnaire. Hence, this information should be retained as part of census dataset.

3. **2 of 7 rule** – this method was developed to detect and remove the false records. 2 of 7 rule states that a response must contain a valid value for at least two of the following seven key variables to be considered to refer to a genuine person:

    - First name or last name in the household section of the questionnaire
    - First name or last name in the person section of the questionnaire
    - First name or last name in the relationship question of the questionnaire (new for 2022)
    - Date of birth requires specifically for the month and year to be populated as a minimum
    - Variables which describe household relationship
    - Sex
    - Marital status

    One of these key variables must be either the name, or date of birth variables. For example, if sex, and marital status were filled in but no other 2 of 7 variables that describe household, this would not pass the 2 of 7 rule.

    There are cases when not all seven variables will exists in a record. However, RFP processing will still require at least two variables, one of which needs to be either name or date of birth. Examples of these cases include:

    - Single person households are exempt from the relationship criteria as there are no household relationships to be recorded. Therefore, the rule becomes 2 of 5.

- Individual questionnaires of people living in communal establishments are exempt from name variable in the household section, and relationship criteria. Therefore, the rule becomes 2 of 4.
- Submitted and unsubmitted online returns, where another household return did not exist – for 2022 online return cannot be submitted without entering both, name and date of birth, automatically should pass 2 of 7.

Name and date of birth are a minimum requirement in data processing, and thus considered when setting the minimum viable return criteria for the online census questionnaire.

4. **Administrative data check**[20] – the RFP process will use administrative data sources to check the plausibility of records where there is insufficient information for the record to pass the 2 of 7 rule, but potentially enough information that a match can be found through administrative data. If the match is found, then the record will be treated as a pass of the 2 of 7 rule. This would provide the confidence to include the record as referring to a genuine person.

Method used in 2011

The 2 of 6 rule was developed to identify and remove the false records for the 2001 Census. Following this method, to be considered a valid record for an individual, the record must contain valid values for two out of six key variables. It remained the only method for RFP through to the 2011 Census.

The additional checking for false names and checking against the administrative data are the new methods for Scotland's Census 2022. As there is an additional

[20] For detailed methodology, see: PMP013: Missing and Different Dates of Birth | Scotland's Census (scotlandscensus.gov.uk)

variable group in the question about the household relationships on the 2022 paper questionnaire, the 2 of 6 rule is now the 2 of 7 rule.

Methods used by ONS and NISRA

ONS are also using RFP process to identify and remove census returns that do not contain sufficient information to be treated as a response. This will include analysis of names and other variables to identify false people not identified in the removal of false persons process.

5.5.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Check for and remove false names**

   To check for false names, firstly the RFP process will identify blatantly false name information. These are the data with a strong indication that the records do not apply to a genuine individual.

   Once these false names have been identified, the resolution of these records can have one of three outcomes:

   1. The record does not belong to a genuine person. These records will be removed in the RFP process if they further fail the 2 of 7 rule.

   2. The name is obscured on purpose. The name variable for these records will be assigned the value 'missing'. This is so that these records can then be passed on to the 2 of 7 rule process, and later to the administrative data matching.

3. Some records will be sent for clerical review. This is in case the false names were captured incorrectly. For example, the scanning of a paper questionnaire might interpret a genuine name 'NIA' as the abbreviation for 'not applicable' – 'N/A'. In this case the clerical review will assess it is a genuine or false name.

Statistics on the overall process will also be recorded, including the number of records changed and the proportion of records removed following the false name check process. Summary statistics frequency tables will also confirm the most common false name or names. Sense checks will be carried out on these to ensure the majority of false names were identified correctly.

For further quality assurance checks, there will be another clerical review on records that failed the name check. For example, false names such as 'NO ONE', or 'NO ONE IS HERE'. These suggest that the record does not contain any personal data. However, the presence of such name is not strong enough evidence for the record to be removed from the census dataset. Hence, these records will still be passed onto the 2 of 7 rule to confirm, and only then removed.

In addition, a similar check will be performed on the records identified as containing a false name. These will be either real names that are incorrectly identified as false by the system (for example, 'NIA'), or names that are obscured on purpose (for example, 'ANON'). Both of these types of records relate to genuine persons and will be passed for further processing, specifically to the 2 of 7 rule process.

2. 2 of 7 rule check

The overall number of records that fail the 2 of 7 rule is not expected to be high due to the built-in validation of the online questionnaire. This will mean that the vast majority of responses submitted online will pass the 2 of 7 rule.

Any records that pass the 2 of 7 rule will be included in the dataset, and processing will continue.

Quality assurance checks include assessing a summary statistics report of records that were accepted and rejected by the RFP process, to ensure that the RFP process is running according to the process methodology. In addition, quality assurance checks include the demographic statistics of accepted or rejected responses by the following characteristics: name, date of birth, sex, marital status and household relationship. Quality assurance checks also include sample statistics of accepted and rejected records based on different geographies to ensure that there is no unusual patterns in the number of accepted or rejected records.

Some of the cases may be classed as borderline by the process algorithm, meaning these do not meet the criteria for automatic acceptance or rejection of the record. All these records will be clerically reviewed.

3. Administrative data check

Any records identified as a marginal pass for the 2 of 7 rule will be referred for matching to an administrative data source to confirm if the record refers to a real person. If a match is found in the administrative dataset, the record will be treated as a pass for the 2 of 7 rule. The matching algorithm is based on a scoring system. Matching will not be possible if there is no name or date of birth in the record.

In terms of scale, the number of records to be referred for matching to administrative data are expected to be in the 1,000s.

There are some cases where a record may fail the 2 of 7 check, but includes some key information (name or date of birth) that would enable the record to be matched to an administrative data source. By coupling this with the postcode information from the associated questionnaire, a corresponding record found in the administrative dataset would provide an indication that the census record represents a genuine person, even though it would normally fail the 2 of 7 check.

5.5.3 Strengths and limitations

Checking for and removing false names and administrative data check are the new RFP methods for Census 2022. These processes allow for a more detailed examination of the data, and increases the likelihood of genuine persons responses are being retained in the census dataset compared to the 2011 Census methodology for RFP. Both additional checks are implemented through statistical algorithms, which reduces the processing required if these were to be carried out manually. Hence, the processing accuracy is increased without a large impact on the timeliness of the process.

In addition, the quality assurance does contain a manual check with a decision-making element of a clerical review. The process in itself is relatively automated involving a reviewer following a set of specific guidelines to reduce potential bias. This is only done on borderline cases where there was not enough information for the automated algorithm to either accept or reject the record. The scale of this clerical review for borderline cases should be minimised by the improvements in the overall methodology and the majority of the census responses completed online, which has built-in functionality to prevent returns that are identified as false persons.

5.5.4 Section summary

Remove False Persons (RFP) process identifies census responses that might not relate to genuine persons and scores them on how likely these are to be false. Where possible, responses that are identified as not representing real people, are not passed on for further data processing to avoid overcounting the number of people in the census.

To quality assure the RFP process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data. These include looking at the profile and proportion of records that are accepted and rejected by the algorithm as false persons, a sense check of

the names that are identified as false, clerical review of borderline cases, and matching of the records to administrative data.

5.6 Data Cleansing – Resolve Multiple Responses

Each household in Scotland must complete and return a census questionnaire for Scotland's Census 2022. People sometimes give multiple census responses in error, and these need to be resolved into a single record in order to avoid overestimating the population. The duplicate records can be present in multiple responses for the same household, individual responses, or responses from individuals living in communal establishments.

The Resolve Multiple Responses (RMR) process in census data cleansing looks to identify and resolve cases where this occurs within the same household or postcode. The process consists of two separate stages: identifying of duplicates and resolving of duplicates. The identified multiple responses are resolved by combining the information on each questionnaire into one response. Including those multiple responses without resolving will result in the overcount of the population of Scotland.[21]

The quality assurance checks for the RMR process include:

- reviewing a summary statistics report on the counts of duplication and multiple cases that have been identified and resolved by the RMR process;
- manual sample check of the links with different matching scores that were either automatically accepted or discarded as duplicate records;
- manual check of the prioritisation rules applied to combined records.

[21] For detailed methodology on RMR identify, see: PMP014: Resolve multiple responses - identify duplicates | Scotland's Census (scotlandscensus.gov.uk);
For detailed methodology on RMR resolve, see: PMP021: Resolve Multiple Responses Prioritisation and Resolution

5.6.1 Background and introduction

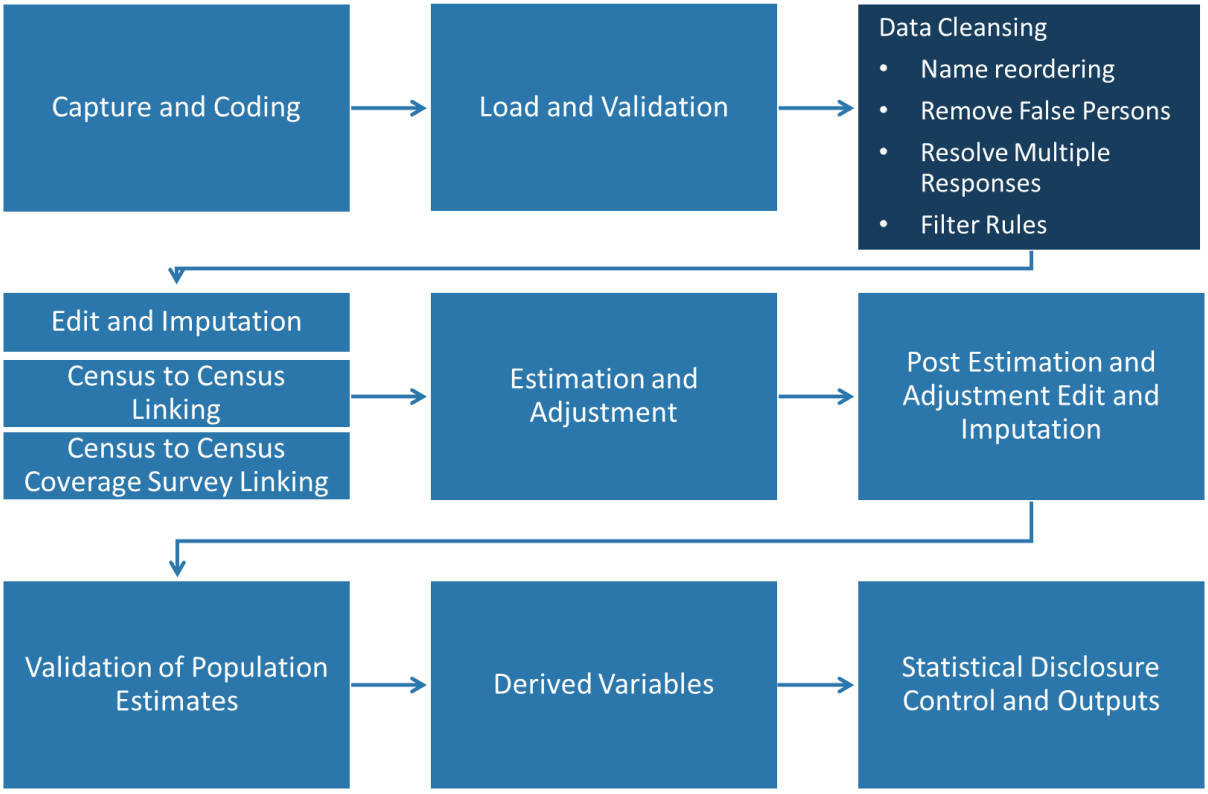
Resolve Multiple Responses (RMR) is a process designed to identify and resolve duplicate records within census dataset. These duplicate records occur when one person or household responds more than once. This can happen for multiple reasons:

- someone in the household completes and submits the census questionnaire, but someone else in the household has already done this;
- a respondent changes their mind about what they want to include in their response and submits a new one;
- a respondent begins filling in the census return online but decides they would rather complete it in on paper. In such cases the information on the online return would be collected as an unsubmitted return;
- a respondent begins filling in the census return online, but forgets their login details before completing it. They would then need to request a new Internet Access Code (IAC) and begin a new return[22]. The information on the first return would be collected as an unsubmitted return;
- a respondent gets confused about their paper response, and answers the individual questions for themselves for Persons 1–5 instead of just Person 1 (each paper household questionnaire contains spaces for up to five people).

It is important to resolve these multiple responses to ensure that there is no inflation of the counts of people or households.

[22] For data security reasons, individuals cannot be given access to partially completed census returns over the phone.

Figure 9. Resolve Multiple Responses process within the census data journey as part of Data Cleansing



The RMR process happens in two stages: identifying and resolving.

1. **Identify stage**:
    - Identify duplicates – records are matched to every record in the postcode to ensure that they are unique in this postcode. Any matches are recorded.
    - Administrative data check – checks to see if a potential duplicate links to multiple administrative data records, then they are likely similar but distinct people.

2. **Resolve stage**: the identified matches are resolved through an automatic process.

During the identify stage, any duplicates within a postcode are identified. This step produces a considerable amount of links of varying strength. To reduce the set of these links, administrative data are used to verify whether these responses refer to

the same or different people or households. This step checks whether two or more people with similar name and date of birth do live in that postcode. Any links that are discovered to be non-matches do not continue to the resolve stage as these are not assumed to be multiple responses and in fact refer to different people.

During the resolve stage, the matches are combined into one record. The process uses a prioritisation mechanism that determines which return is the primary one. Priority is given first on response type (online return over paper return over unsubmitted online return). This is because the online return is more likely to be completed fully and accurately due to built-in functionality that validates the responses at the time of completion. The online questionnaire will prompt a respondent if they enter a non-plausible or conflicting value. For example, the prompt will appear if someone enters a date of birth suggesting they are 200 years old, which can be easy corrected by the respondent. These types of mistakes are more likely to be unnoticed by respondents completing paper questionnaire. If there is more than one response of the same type (for example, two paper returns), the most complete response is prioritised. In the case of a matched record from one individual questionnaire and one household questionnaire, the individual questionnaire takes priority.[23] All the information present in the primary record is retained, where there is missing information it is taken from the secondary return, and so on.

One important exception is a non-response on voluntary questions. Some questions in the 2022 census are voluntary so a non-response is in effect a valid response to these questions, rather than a missing value. In this case, the non-response from the individual questionnaire will not be replaced by a value from the household questionnaire.

[23] See more details in the section on individual questionnaires here: Questionnaire completion | Scotland's Census (scotlandscensus.gov.uk)

Methods used in 2011

The RMR process was the main method of identification and resolving duplicate records in 2011 Scotland's Census. However, in 2011, the way matches were identified differed slightly due to different enumeration strategies. In addition, the 2011 RMR process did not use administrative data to verify matches.

Methods used by ONS and NISRA

ONS are using the analysis of the linkages of the census returns with other data sources to identify and remove returns that relate to a person already correctly covered in another return in the same location. ONS tested this approach by linking the 2011 census data to the Longitudinal Study.

5.6.2 Proposed method for quality assurance in 2022

As described above, RMR process is applied in two stages: identifying of duplicates and then resolving duplicates. The following quality assurance checks will be conducted to ensure the process is working correctly.

Identify stage

1. **Identify stage — summary statistics reports**

   A summary statistics report will detail the counts of duplication and multiple cases that have been identified by the RMR process. The report will also include the overview by categories of matches of each link by the matching algorithm. These categories based on a range of strength scores along with additional evidence of possible duplication case, like the information about the household relationships[24].

[24] For detailed methodology on RMR identify, see: PMP014: Resolve multiple responses - identify duplicates | Scotland's Census (scotlandscensus.gov.uk)

The duplication rate will also be analysed by comparing counts of duplicate returns with the total number of returns. If the number of returns is unexpectedly high, a higher rate of duplication is expected. This may happen if a large number of people forget their password for the online questionnaire as this will result in duplicate returns since the original questionnaire will also be submitted as an unsubmitted online return. For example, there might be two returns from the same persons if they started but did not finish completing their online census questionnaire as they forgot their password. And another return that they completed after requesting a new Internet Access Code (IAC). In this situation, there will be high return count (two returns for this person) as well as high duplication count (two duplicate returns). In other words, the high duplication count is explained by the high return count.

However, if the amount of responses is high, but the amount of duplication is low, this may indicate a problem in enumeration of households or in the duplication identifying process. A high duplication count should be consistent with a high return count, and vice versa. Any deviation from this pattern will be investigated further.

A summary statistics report will also show the counts of identified cases of duplication along with the scores from the matching algorithm. These statistics will be stratified by form type (paper, online, individual, household, and so on) and certain individual characteristics. This will inform the whether there are any differences in types of questionnaires or respondents' characteristics that caused people to submit multiple responses.

2. Identify stage — sample check

A sample check of the links that were assigned a strong matching score and were therefore automatically accepted as duplicate records will be carried out. This check will be completed early on in the process to identify any systematic errors. To ensure consistency of the process, the check will be completed on various types of returns (including online responses, paper responses and online

unsubmitted responses). Where possible, the check will be repeated at strategic points in the process; for example, when the first of each type of returns is received in order to identify any systematic problems earlier on.

3. Identify stage — clerical review

A clerical review will be carried out on a sample of the matches, which were not strong enough to be automatically identified whether they are duplicates.

During a clerical review, a reviewer assigns relevant values to the records to be either accepted or rejected as duplicates. The process is based on specified set of rules, however involves a level of decision-making.

A sample of items sent to clerical review will be taken and the clerical review process will be repeated by a different .reviewer. The results will then be compared to the outcomes of the primary clerical review. This is to ensure the consistency in the decision-making. The agreement rate of both results will demonstrate the consistency of the overall clerical review process.

In addition, if necessary, internal statistical teams will provide statistical peer review of any changes to clerical review business rules made in reaction to issues found during census live operations.

Resolve stage

1. Resolve stage — summary statistics report

A summary statistics report will be produced on the counts of resolved records for households, persons completing an individual census response and persons living in communal establishments (CEs)[25] . It is expected that the count will

[25] A communal establishment is typically a managed residential accommodation where there is full-time or part-time supervision of the accommodation. People living in communal establishments must complete and submit an individual census response. See Glossary for more details.

match the number of person duplicates or multiple responses from the identify stage.

2. Resolve stage — manual sample check

When identified duplicate responses are resolved into one record, the process follows a specific set of prioritisation rules in order to use one response as a primary record, and other responses as secondary records in case of any missing information in the primary one. To ensure the automatic prioritisation procedure assigns the records correctly, a manual check will be carried out on a sample of these records. A reviewer will manually compare the algorithm's decision to the prioritisation rules for the records in the sample. The sample will be relatively small as this is a check of pre-determined algorithm. The focus will be on including different types of returns (paper returns, online returns, and unsubmitted online returns) in the sample to ensure the algorithm works for all of these correctly.

3. Resolve stage — voluntary questions check

According to the RMR methodology, in cases of record duplication from different types of response, the information in the household response will be removed in favour of the individual response. This is because it is assumed that the individual response might be submitted independently from the household response and will be of a higher quality. Because of the prioritisation of the individual responses, the same process will be applied to voluntary questions where a non-response on the individual response will be chosen over a different response on the household response.

Hence, a separate check will be carried out to ensure this process is applied correctly during Census 2022. This check will include samples that contain different types of returns (paper returns, online returns, and unsubmitted online returns). The check will be performed by comparing the records that have been

combined into one and checking that a non-response in an individual questionnaire is not replaced by a response in a household questionnaire.

In addition to this, the check will examine the information that was not retained. For example, if an individual questionnaire is prioritised, the information in the household response, and thus being removed, will be examined.

4. Resolve stage — distributions report

A report on the distributions of responses for all questions for before and after RMR datasets will be produced. Analyses of these distributions will also be done by the characteristics of respondents who submitted multiple responses. This will not inform any immediate changes to the process, and instead will be used as part of evaluation.

5.6.3 Strengths and limitations

The majority of responses in Scotland's Census 2022 will be online. The functionality of the online census questionnaire allows for an easy completion by the respondents. However, there is also an increased potential that people might request additional Internet Access Codes (IACs) because they have forgotten or did not set a password for their online questionnaire, and, thus, are not able to complete an already started response. This is due to the security reasons for protecting personal data. This could cause a larger number on partially completed online returns that will in turn result in duplicate records when they are received as unsubmitted online returns. However, the automated RMR processes are designed to identify and resolve duplicates efficiently, and a potential increase in volume of these duplicates should not greatly affect the processing effectiveness or timeliness.

5.6.4 Section summary

In some cases, a person or a household will complete and submit more than one census response. This could be through an online or paper questionnaire, or both. As part of Data Cleansing stage of the census data processing, the process of Resolve Multiple Responses (RMR) is designed to identify and resolve these duplicate records. The resolving of the identified duplicates involves combining these into one most complete record.

To quality assure the RMR process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data. These include a number of summary statistics reports to inform the progress and identify any issues, manual sample checks to ensure the correct running of the automatic processes, and a number of additional analysis for evaluation to inform future development.

5.7 Data Cleansing – Filter Rules

Every household in Scotland must complete a census questionnaire. However, not everyone has to answer every question in the census. This is because some questions are not relevant to specific respondents due to, for example, their age. In these cases the questionnaire guidance will direct the respondents to what question should be answered next or which questions to skip. This will be applied automatically in the online questionnaire. For paper questionnaires the respondents will follow the instructions in the questions.

Sometimes respondents may not follow the guidance, and answer the questions that are not relevant to them. Filter Rules process identifies and resolves these inconsistencies in the data.

Filter Rules is a process that identifies and where possible resolves issues in the answering path (routing) of a questionnaire. In 2022, this will mostly be required for paper questionnaires. The online questionnaire has a built-in routing function that will automatically direct respondents to the relevant questions. This functionality should greatly reduce the number of routing errors for online responses.

The Filter Rules process identifies and where possible resolves any issues where the questionnaire routing was not followed. In some cases where the overall response contains enough information for the errors to be corrected by the process, the incorrect or inconsistent responses will be assigned a new value. For example, if someone answered ‘No, I have never worked’ in the Ever Worked question, but provided full details on their employment in further questions, it is reasonable to assume that the first answer was made by mistake. In this case, there is a degree of confidence for the mistaken value in this the record to be changed. However, if there is not enough information to make this decision with certainty, the original value in the record will kept and be passed on to the Edit and Imputation process that will correct this inconsistency using a different method (see Section 5.8).

The quality assurance checks for the Filter Rules process include:

- assessing the performance of the overall process using the overall counts of the records identified for the application of filter rules;
- ensuring that the final dataset after the process does not contain any incorrect record combinations;
- manually checking a sample to ensure the records have been assigned correct values following the process.

5.7.1 Background and introduction

Filter Rules is a process that identifies and resolves issues in the answering path (routing) of a questionnaire. In Scotland's Census 2022, this will be required mostly for paper questionnaires.

For the paper questionnaire specifically all blank questions are automatically marked as 'Missing' during the Coding stage. Filter Rules process checks which related questions have been answered. If the questions are blank because of the appropriate routing, in other words, the correct answering path was followed, the Filter Rules process changes these values to 'No code required'. Instances where the routing rules were not followed these will be resolved at the Edit and Imputation stage of the Data Cleansing process.

For example, the question about tenure (Figure 10) has routing instructions to the question about landlord (Figure 11). The instructions within the response options guide respondents to answer the next question or skip it depending on the response option they chose. If response options 'Owns with a mortgage or loan', 'Owns outright', 'Owns with shared equity', 'Part owns and part rents' are chosen, the question about the landlord should be skipped as it does not apply to this circumstance. If the respondent correctly leaves the landlord question blank, the value of 'Missing' is changed by the Filter Rules to 'No code required'.

This routing is applied automatically in the online questionnaire.

Figure 10. Question about tenure on paper household questionnaire (paper questionnaire)

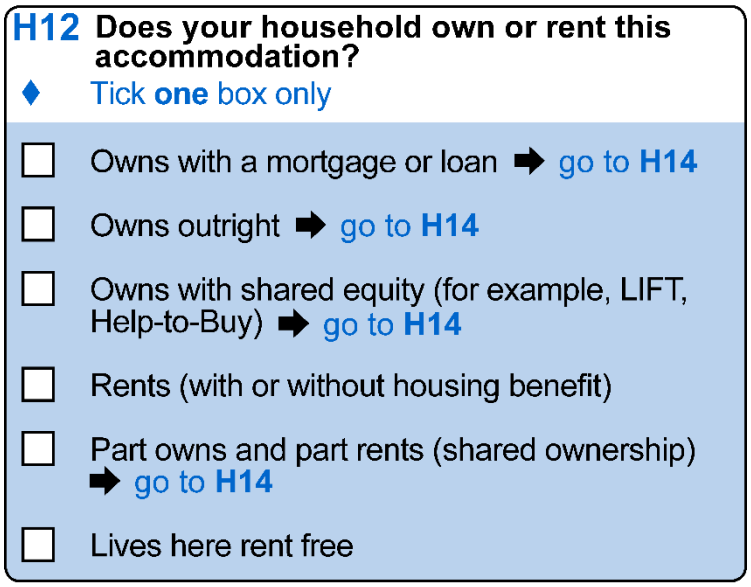
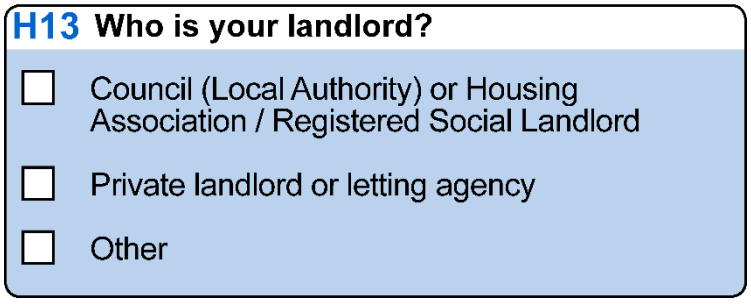
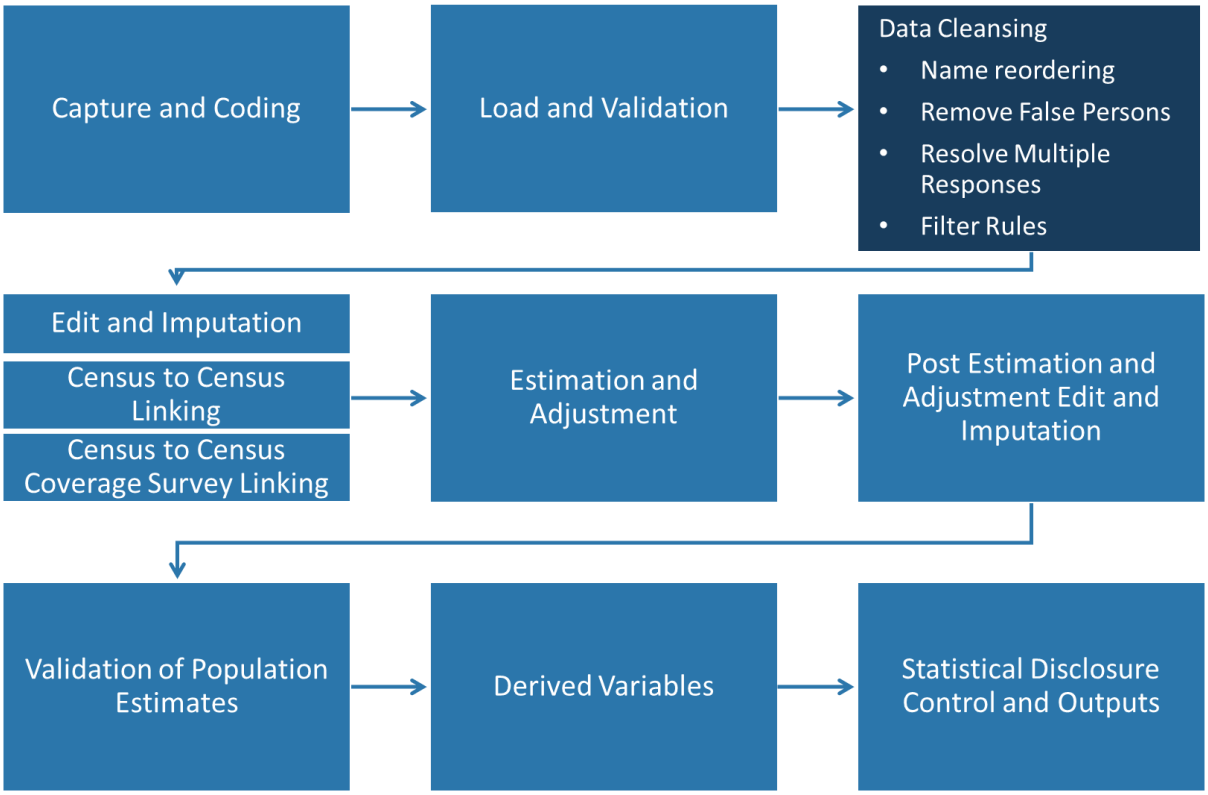


Figure 11. Question about landlord on paper household questionnaire (paper questionnaire)



However, it is possible that the guidance is not followed and the respondent does not follow the correct answering path and answers the question about landlord. In addition, respondents completing paper questionnaires might cross through the questions that do not apply to them, which can be picked up during the scanning and capture process as a response if the line is drawn through a tick box. Similarly, the routing rules are not followed if the respondent should have answered the landlord question, but incorrectly missed the question. This will result in an incorrect or inconsistent value for that response. All responses with inconsistent values will be indicated as such, and will be passed on to the Edit and Imputation process for further processing.

Figure 12. Filter Rules process within the census data journey as part of Data Cleansing



Questionnaire routing is applied to the following questions[26]:

- Tenure and Landlord
- Student Status and Term-time Address
- Country of Birth and Date of Arrival
- Currently working
- Ever worked
- Workplace Address and Method of Travel to Work

Questionnaire routing based on a respondent's age is applied to the following questions:

[26] For a full question set that contains the final questions for Scotland's Census 2022 see: Question set | Scotland's Census (scotlandscensus.gov.uk)

- Trans Status and History
- Marital Status
- Sexual Orientation
- Address One Year Ago
- Unpaid Care
- Language questions
- Qualifications
- Ex-service status
- Labour market questions

The routing rules in the census questionnaire are formed based on a thorough development and are intended to reduce and correct various possible errors in the answering path. However, it is possible that the data will reveal an unexpected pattern during live operations. The overall quality assurance strategy includes the acknowledgement of this possibility and planning of relevant options.

Methods used in 2011

The Filter Rules process was used in 2011. Many of the filter rules have been retained for Scotland's Census 2022, especially where the questions are unchanged. Improvements have been made since 2011, in particular, the processing of these filter rules now takes place within the Canadian Census Edit and Imputation System (CANCEIS)[27].

Scotland's Census 2022 will be conducted primarily online, with the online questionnaire as the main method of census completion. The online questionnaire includes an additional functionality of built-in routing and validation. This will aid the

[27] Canadian Census Edit and Imputation System (CANCEIS): software designed by Statistics Canada and used during the edit and imputation process. Contact canceis@canada.ca for more information on CANCEIS.

respondents in completion of the questionnaire and increase the quality of data coded at the time of completion.

A number of questionnaire routing rules were introduced for Scotland's Census 2022 as a result of new census questions. In addition, a number of filter rules were modified based on the evaluation of respondent behaviour from 2011. For example, the online questionnaire includes an additional routing for language questions, so that it is not required to answer these questions for persons under the age of three.

Methods used by ONS and NISRA

The Filter Rules process involves applying logic in identifying and where possible resolving errors in the answering path. The deterministic rules for this process are created following the answering path of the census questionnaire. This principle of filter rule application is the same across the UK census offices. However, individual filter rules differ as the questions and question routing for these questions are different across the questionnaires in Scotland, England and Wales, and Northern Ireland.

5.7.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Summary statistics**

   Summary statistics reports will allow analysis of the number of records changed by the Filter Rules process. This will include the number of the records assigned a new code ('No code required') within the dataset. It is unlikely that the Filter Rules process will assign 'Missing' values. However, should the process assign any other values other than 'No code required', the summary statistics report will detail the number of records changed. These summaries will also show whether there are any possible incorrect record combinations still present in the census

dataset after the Filter Rule application. The report will be reviewed and evaluated, and details of any subsequent conclusions or actions will be recorded.

The correct application of the process should result in the values to be assigned either 'No code required', or left as the original values. In cases where the inconsistent responses are left as the original values, the records will be dealt with by the Edit and Imputation process. Further quality assurance checks will be conducted during the Edit and Imputation process (see Section 5.8).

This check will help to identify whether there are potentially any unusual patterns in the census data.

2. Sample check

A manual sample check of the records that have been assigned a new value following the Filter Rules process will be carried out to ensure the algorithm assigned these values correctly. This check will involve comparing the data before and after the Filter Rule process by examining individual records within that sample.

This check will complement the information on any unusual patterns in the census data identified in the summary statistics check.

5.7.3 Strengths and limitations

The Filter Rules process is deterministic, however, to ensure that the process is running in accordance with the pre-determined methodology, a number of quality assurance checks have been developed.

The digital-first approach for Scotland's Census 2022 will encourage the majority of the population to complete their census online. The functionality of the online questionnaire allows the application of the filter rules and validation of the responses at the time of completion. Thus, the requirement for the Filter Rules process is likely

to be reduced for online responses. The majority of the census records requiring this correction will be limited to a lower volume of census responses from paper questionnaires.

As mentioned previously, there is a possibility of an unusual and unforeseen respondent behaviour during the census. This might require additional changes, such as implementing additional filter rules to the existing methodology. A group of Census statisticians will review any changes, and any new filter rules will undergo the same quality assurance as described in this section, including summary statistics and a manual sample check.

5.7.4 Section summary

The Filter Rules process identifies and where possible resolves issues resulting from inconsistencies in the answering path (routing) of a questionnaire. This occurs when respondents answer questions that they were meant to skip, or skip questions that they were meant to answer. The Filter Rules process also identifies inconsistent information that is found in answers to groups of questions, which can also cause issues with later statistical processing. These inconsistencies are corrected taking into account other answers in the questionnaire.

To quality assure the Filter Rules process during live operations for Scotland's Census 2022, a number of statistical quality assurance checks will be performed on the data. These checks include reviewing the summary statistics throughout the process to identify whether filter rules have been applied to a larger than expected number of records, and identifying any unusual or unexpected patterns in the data resulting from respondent behaviour. Samples of individual records will also be checked to ensure that the routing has been applied according to the process methodology.

Some of these quality assurance checks are resource intensive, especially if additional filter rules are required in addition to those planned in advance.

5.8 Edit and Imputation

Scotland's Census 2022 asks every person in the country questions about themselves and the people they live with. Respondents must answer every question, and can choose to answer those labelled as voluntary, and do not have to answer any of the questions that the respondent is instructed to skip because the questions are not relevant to them (this is an automatic process for online respondents). For example, the question about marital status will not be asked of people under the age of 16.

Despite every effort to help and encourage respondents to fill out the questionnaire as accurately and completely as they can, there will inevitably be records which have missing responses to some questions. There will also be some respondents who make mistakes when answering questions, which can lead to inconsistencies across a return. For example, if someone writes the current year instead of their birth year for the Date of Birth question, they will appear to be zero years old and yet may be married, have qualifications, a job, and so on.

The Edit and Imputation process identifies the missing and inconsistent responses and uses a method of donor imputation to fix these by replacing them with valid values from a donor record, which is a similar record to the original in structure and geographic location.

The quality assurance ensures that the process is run according to the Edit and Imputation methodology[28]. This will include the following quality assurance checks:

- assessing the performance of the overall process using the diagnostic reports of the proportion of missing values, inconsistent values and outliers;
- manually checking a sample of records to ensure the edit rules have been applied correctly.

[28] For detailed methodology, see: PMP012: Overview of edit and imputation for Scotland's Census 2022 | Scotland's Census (scotlandscensus.gov.uk)

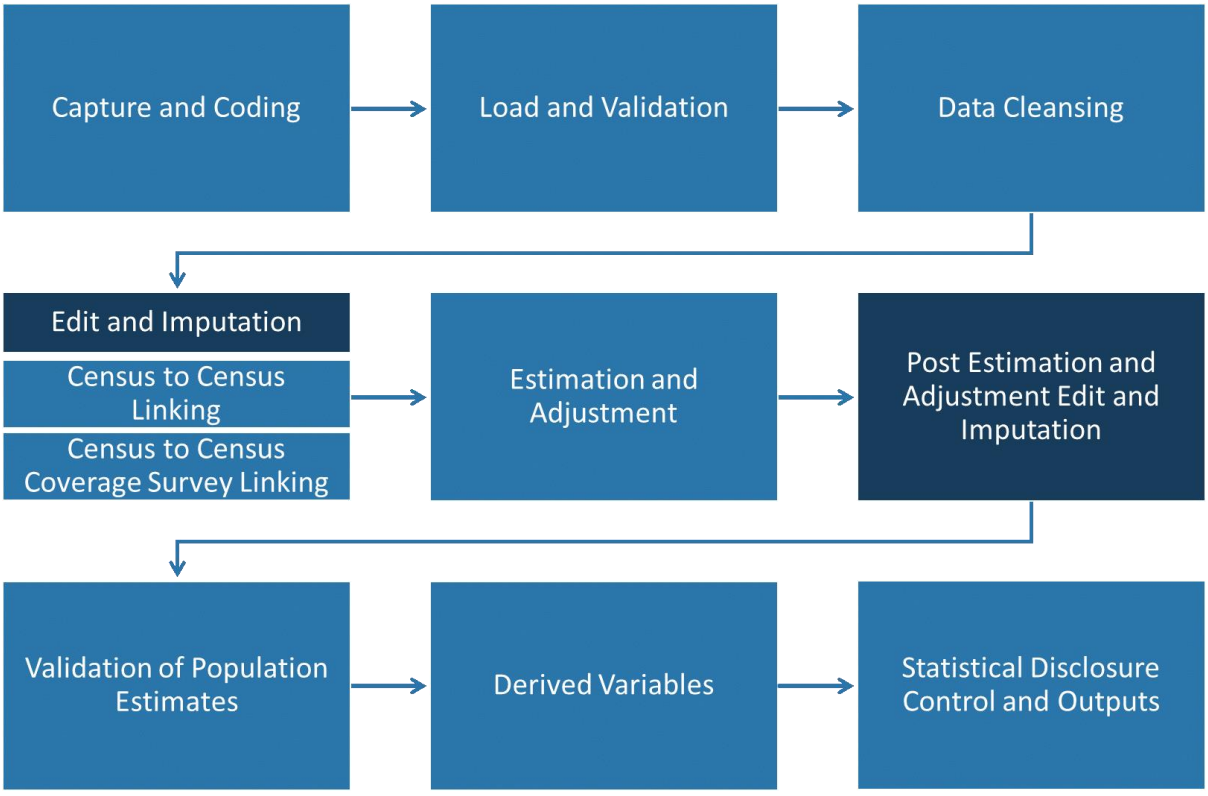
5.8.1 Background and introduction

Although every effort is made to collect full and accurate census responses, inevitably there will be some incomplete census returns due to respondents not answering all required questions. Census returns may also contain invalid and inconsistent responses, which may be due to respondent error or procedural error, such as limitations of the data capture process for paper questionnaires. There is an expectation that the main census outputs are complete and consistent. Hence, Edit and Imputation process deals with these incomplete and inconsistent records before the production of census outputs.

The Edit and Imputation process identifies these missing and inconsistent responses, fills in any blanks and corrects inconsistencies using robust statistical methods to produce plausible results. This process cannot predict with absolute certainty that the value assigned by the process is true for an individual, but the overall effect will be that the outputs produced from census data will more accurately reflect the population than had the process not been carried out.

The main method used in Edit and Imputation is called donor imputation. For each inconsistent or missing record that needs to be fixed, the process looks for similar records in the census dataset and then incorporates responses from the donor record, in order to fill in the blanks or correct inconsistent responses.

Figure 13. Edit and Imputation process within the census data journey



The main Edit and Imputation process is applied to the census data after the Data Cleansing process. The entire Scotland dataset is processed at this stage. This allows to improve imputation process and data quality as a whole, especially for imputation of large households.

The process of Edit and Imputation is then applied again in a similar way to the census data after the Estimation and Adjustment process. This is known as Post Estimation and Adjustment Edit and Imputation. See the section on Estimation and Adjustment (see Section 5.11) for more details.

The Edit and Imputation process uses robust statistical methods to produce plausible results. The overall process consist of two main elements:

1. imputation of household relationships – this focuses on correcting any inconsistencies in the responses to questions on how people in a household are related to one another;

2. edit and imputation rules for individual questions – this is used to correct the missing and inconsistent responses in an individual record for a person.

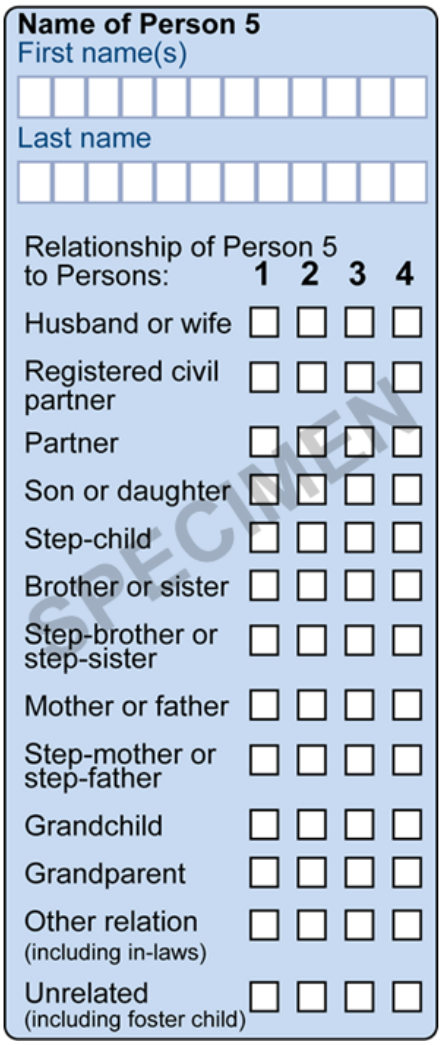
The household relationships element of the process is very complicated and, hence, applied to the data separately before dealing with the imputation of the individual questions responses. The sections below give a brief overview of these processes[29].

1. Imputation of household relationships

Some errors occur during completion of the question about how the members of the household are related to one another. It should be noted that these errors in completion mostly occur when respondents complete a paper census questionnaire. See Figure 15 for an example of the relationship question on paper questionnaire for person 5 in the household. The online questionnaire has built-in functionality that aids respondents in completing this portion of the census, which will reduce the number of errors.

[29] For detailed methodology, see: PMP012: Overview of edit and imputation for Scotland's Census 2022 | Scotland's Census (scotlandscensus.gov.uk)

Figure 14. Example of Scotland's Census 2022 household relationship question for Person 5 (paper questionnaire)



Based on the processing experience from 2011 census, common errors in the responses for the household relationship question are likely to include:

- getting relationships within the household the wrong way round;
- ticking the relationship to the person filling in the questionnaire rather than the relationship to the person indicated in the column;
- not filling in the persons in the same order consistently throughout the household questionnaire, so that age, sex and marital status do not align with the recorded relationships.

In order to account for some of these inconsistencies, Relationship Algorithm 1 (RA1) has been developed. It is a deterministic algorithm applied using CANCEIS[30] software prior to donor imputation. RA1 corrects the most common respondent errors. This includes listing relationships the wrong way round (for example, someone intending to tick a response option as "I am his parent" but actually ticking the response option "He is my parent") and missing relationships (three-generation relationships, missing siblings or parents). Since this is a deterministic process, the algorithm uses very strict criteria to identify records which it will amend. For example, when looking at a parent-child relationship where the child is older than the parent, the following criteria must all be met before the parent-child relationship is reversed:

- Younger person is (mis)reported as parent of older person
- The age gap between the two people is at least 13 years
- The older person is at least 16 years old
- The younger person is no more than 30 years old
- The younger person has a marital status 'single'
- The younger person has no partner, spouse or civil partner in household.

Testing on 2011 data shows that RA1 significantly reduces the number of households requiring imputation and increases the quality of imputation and chances of successful imputation by donor. Any relationships not fixed by the RA1 are fixed in the subsequent processes.

2. Edit Rules

Edit and Imputation processes the data in modules, which are collections of variables that are to be imputed at the same time. This is done in order to improve the imputation, so that some inconsistencies can be resolved by

[30] Canadian Census Edit and Imputation System (CANCEIS): software designed by Statistics Canada and used during the edit and imputation process. Contact canceis@canada.ca for more information on CANCEIS.

imputing variables at the same time, rather than being dependent on which of those variables imputed first. For the modules, the variables are grouped thematically, so that these contain variables that relate to one another and are predictive of each other. Table 3 below lists the variables for individual questions grouped within each imputation module[31].

Table 3. Variables within the Edit and Imputation modules

Demographics	Culture	Health	Labour market
• Age • Sex • Marital status • Full-time student • Term-time location • Relationships • Economic activity	• Address 1 year ago • Country of birth • Date arrived in UK • Ethnicity • National identity • Language questions	• Carer • Disability • Long-term conditions	• Qualifications • Ever worked • Hours worked • Employee status • Supervisor • Industry • Occupation • Work/study address • Method of travel

Within each module, the Edit and Imputation process includes a number of edit rules, which identify and change any inconsistencies or outliers. These are known as hard edits and soft edits:

- Hard Edits

A hard edit specifies situations that the Edit and Imputation process does not allow in the dataset. These are something that is impossible or so rare that most occurrences are errors. For example, a person under the age of 17 cannot drive to their place of work or study.

[31] For detailed methodology, see: PMP012: Overview of edit and imputation for Scotland's Census 2022 | Scotland's Census (scotlandscensus.gov.uk)

- Soft Edits

A soft edit specifies situations that are very unlikely, which are intended to be kept in the dataset but they should not be created disproportionately through imputation. In other words, these records should not be used as donors in the imputation process thus increasing the number of times this unlikely situation appears in the dataset. For example, a person is unlikely to be more than 65 years older than their child is.

Method used in 2011

Scotland's Census 2011 used the Canadian Census Edit and Imputation System (CANCEIS)[32] for the first time. The Edit and Imputation process was implemented using CANCEIS and SAS[33] software. The CANCEIS modules and the set of SAS code were developed by ONS. The code was then adapted for the Scotland's Census 2011 by modifying the input information in collaboration with ONS and Statistics Canada.

In the Scotland's Census 2011 there were three relationship algorithms as part of the Edit and Imputation process:

- Relationship Algorithm 1 (RA1);
- Relationship Algorithm 2 (RA2), which was a repeat of RA1 but was performed after donor imputation of relationships;
- Relationship Algorithm 3 (RA3), which imputed relationships in households which were too large to be imputed as donors.

[32] Contact canceis@canada.ca for more information on CANCEIS.
[33] SAS (Statistical Analysis System) software is widely used for data management and statistical analysis. Refer to the SAS website for more information: SAS: Analytics, Artificial Intelligence and Data Management | SAS UK

RA2 was deemed unnecessary during live running in 2011 as it did not add additional value, and RA3 has been replaced by donor imputation for Scotland's Census 2022.

In addition to applying the relationship algorithm using the CANCEIS software, some complex or unusual household compositions were imputed manually. The manual imputation to correct inconsistencies within recorded household relationships were applied to 3,500 households[34].

Methods used by ONS and NISRA

Similarly to NRS, ONS and NISRA intend to use CANCEIS for the Edit and Imputation process.

5.8.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Review summaries**

   CANCEIS software for the Edit and Imputation process will produce a number of diagnostic reports, including key summary statistics on Relationship Algorithm 1 (RA1) and donor imputation process including:

   - Relationship Algorithm 1 (RA1) summary – a summary report will be produced for each edit rule of the RA1. The report will include the counts of how many times each relationship edit rule has been applied. Data from the 2011 Census will be used to indicate the expected proportion of records that each edit rule will be applied to. These figures will highlight

[34] For a report on household relationships data in Scotland's Census 2011, see: Release 2C - Household Relationships Data Quality Issues | Scotland's Census (scotlandscensus.gov.uk)

any unusual patterns in the data and any large differences between 2011 and 2022 data for further investigation.

- Donor imputation summary – a summary report will include the proportions of missing values, inconsistent values and outliers identified in the data. These will be produced for each module with the edit and imputations themes where a collection of variables are imputed at the same time: demographics, culture, health and labour market.

2. Sample check

A number of missing, inconsistent or invalid values will be imputed by CANCEIS during the imputation process. A sample of imputed variables will be manually checked to ensure that the process applied the edit rules according to the Edit and Imputation methodology. This will also check whether the application of edit rules during the imputation process produced implausible combinations within the records, particularly with respect to household relationships. This will be done for all imputed variables. To manage the overall number of imputed records, a suitable size sample will be chosen for this quality assurance.

A separate sample check will be carried out for the household relationship data to ensure the deterministic algorithm is performing according to the set methodology.

3. Distribution check

The distributions for each imputed variable will be compared before and after the imputation process to ensure these do not have any unusual patterns or anomalies. Should any anomalies be found, these will be investigated further.

The distributions of imputed variables will be checked at Scotland level as well as by Council Area to ensure no geographical bias is introduced by the imputation process. For example, a cross-distribution of imputed data by single year of age

by sex per each Council Area. In addition, where available, distributions for imputed variables will be checked against comparator data sources to ensure these generally align with the existing data, or where there are differences, these can be explained. For example, this will include a comparison with NRS mid-year population estimates, Scottish Government Pupil Census, and data from Scottish Survey Core Questions for the adult population.

4. Communal Establishment records

The responses collected from respondents living in communal establishments (CE) are imputed separately to responses from people living in households. Additional checks will be applied to CE records specifically. These checks will include distributions of responses to each question, checking occupation against the donor type to differentiate between CE staff and CE residents, and checking type of residents against the type of CE (for example, boarding schools, care homes).

5.8.3 Strengths and limitations

A number of quality assurance checks are resource intensive as these will require a manual checking of record-level data. However, this issue can be mitigated by applying the check to a sample of a suitable size.

The digital first approach for Scotland's Census 2022 will encourage the majority of the population to complete their census online. The online questionnaire guides respondents through the completion of the household relationships question. It also includes a number of built-in validation rules and messages to minimise the number of inconsistent, invalid or missing responses. This functionality together with the expected increase in online responses will provide higher quality data being submitted into the imputation process than in 2011, which in turn increases the quality of imputation.

5.8.4 **Section summary**

To quality assure the Edit and Imputation process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.

These include sample checks to ensure the imputed variables adhere to the methodology. There will be a review of summary statistics and the distributions of imputed variables to reveal any unusual data patterns due to the potential errors in the process. These will also be assessed by using a comparator data.

Some of these quality assurance checks are resource intensive, however, can be mitigated by further development of the detailed methods in advance.

5.9 Census to Census Linking

Each household in Scotland is required to provide a census return. The household return includes all people usually resident at that address. Any people usually living somewhere else but staying at that address on the night of the census should be recorded as visitors and not as usual residents for that address. However, sometimes people are recorded as usually living at multiple addresses, or at a wrong address. Including these records as separate distinct responses will result in overestimating the population of Scotland. The Census to Census Linking process is applied to these records to determine whether these are duplicates or refer to different people. The process identifies the duplicate records and then links those to an administrative data source. This information is then used to estimate and correct the total count of population in the Estimation and Adjustment process (see Section 5.11). This section describes the statistical methods that will be used to quality assure the Census to Census Linking process for Scotland's Census 2022.

The quality assurance of this process focuses on ensuring that the linking is performed according to the Census to Census Linking and Overcount Correction methodology[35]. The quality assurance checks will include:

- manual sample checks of linked records to ensure the linking process was performed correctly in line with the methodology;
- manual sample checks of the main process algorithm;
- evaluation of the overall process by assessing the distributions of probabilities.

[35] For detailed methodology see: PMP015: Census to census linking and overcount correction | Scotland's Census (scotlandscensus.gov.uk)

5.9.1 Background and introduction

Census to Census Linking is a process used to account for overcoverage that arises from census records duplicates. The process identifies potential duplicate records and then links them to an administrative data source to determine whether these records represent the same person recorded in multiple locations, or distinct individuals.

There are different types of duplicates that may occur in the census data from people being recorded at multiple locations or the wrong location. There are four types of population overcount:

- Type 1 — Duplication of individuals within the same location[36]
    - These are duplicates where a person has either been included multiple times in the same household census return, or in two or more separate returns for the same household.

- Type 2 — Individuals enumerated in more than one location
    - These are duplicates where a person has been included in more than one household census return at different addresses, such as a child with parents who live apart is included in the household of each parent.

- Type 3 — Individuals enumerated in the wrong location[37]
    - These are cases where a person has been missed in the household where they should have been enumerated, but included in a household where they should not have been enumerated. This results in undercount in the area where they were missed, and overcount in the area where they were included.

[36] Type 1 overcount is dealt with at the RMR process (Resolve Multiple Responses, Section 5.5).
[37] Type 3 overcount is dealt with at the Census to CCS matching process (Census Coverage Survey, Section 5.10)

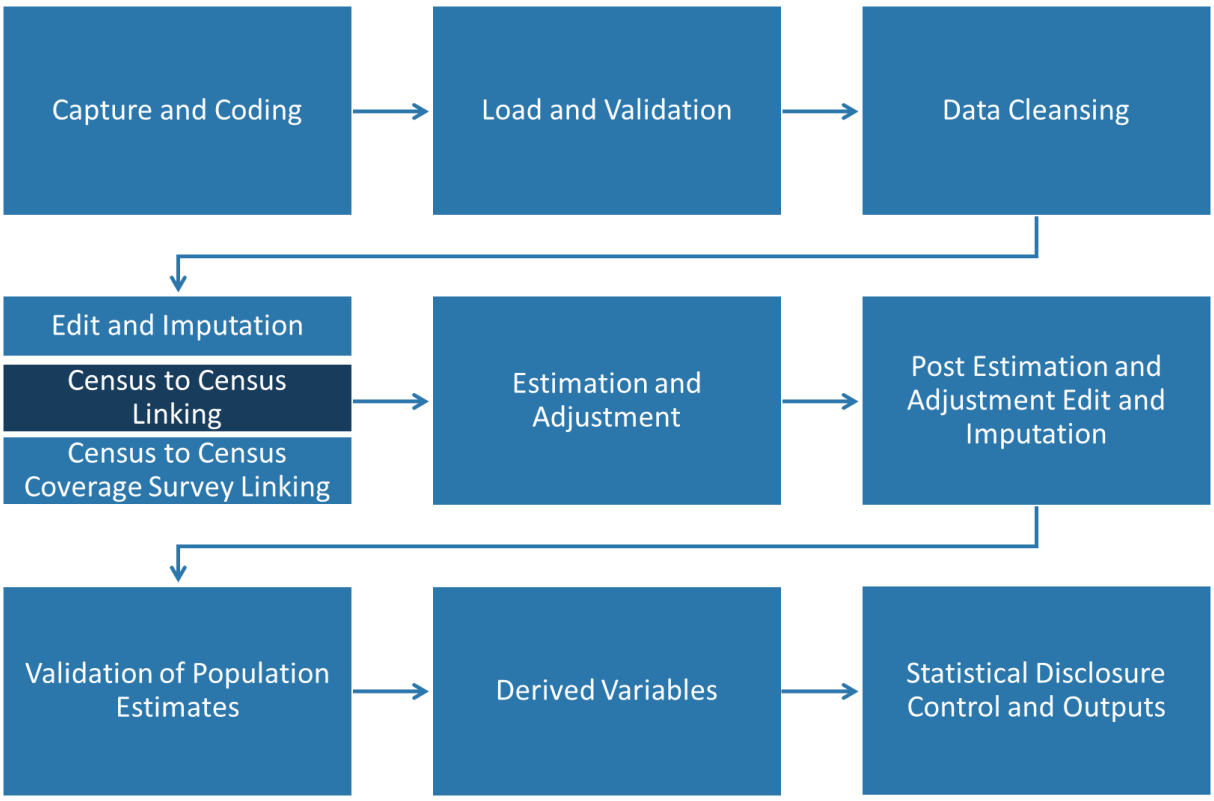
- Type 4 — Erroneous returns[38]
    - These can be returns that are fictitious or joke returns, as well as cases of babies that were born after the Census Day, or individuals who died before the Census Day, and as such should not have been included.
    - These are difficult to identify without additional field work or linking to vital events data (births and deaths registration data).

Census to Census Linking process is applied to account for Type 2 overcount. The process uses a method of linking different census records together to identify any duplicates, which are then linked to an administrative data source. The records are linked by comparing the name and date of birth. Scores are then calculated for each set of linked records to derive the probability of each linked census record representing a distinct genuine individual.

The process of Census to Census Linking takes place after the Edit and Imputation process (see Section 5.8) in the census data journey. The process creates a dataset to be used in the Estimation and Adjustment process (see Section 5.11) to estimate the scale of adjustments needed to correct for any overcount of the population.

[38] Type 4 overcount is dealt with at the RFP process (Remove False Persons, section 4.5).

Figure 15. Census to Census Linking process with the census data journey



In the main process of Census to Census Linking after the initial linking, records are then grouped in bands (or strata) based on:

- strength of agreement on name;
- strength of date of birth agreement;
- how rare the identifiable information is in the population.

How rare the identifiable information is based on how many times the combination of name and date of birth appears in the dataset. For example, it is more likely that two records with name 'John Smith' will appear in the dataset more often than more uncommon names. Hence, it is more difficult to have confidence that these records are a match for the Census to Census Linking. In comparison, two records for the name 'Sarah-Jane Watt-Maxwell' are more likely to refer to the same person and, therefore, are a genuine match.

For a linked pair of census records (in other words, potential duplicates), trying to link each of them to the administrative dataset leads to one of the three possibilities:

0. Neither census record links to the administrative dataset
1. Exactly one census record links to the administrative dataset
2. Both census records link to the administrative dataset

The link between the two census records either represents a match (that is, the two linked records represent the same individual) or a non-match (that is, the two linked records represent different individuals). It is assumed that if a census record links to the administrative dataset then it represents a genuine person in that location. Using these counts, the probability that each census to census link represents a match can be calculated. From that, the probability that each census record represents a distinct genuine individual can be calculated. Comparing the difference between the total number of census records, and the sum of these probabilities, gives an estimate of the total overcount in the census dataset.

For example, if exactly one of the census records is linked to administrative data, this census record will be assigned a probability of one (1). In other words, it is certain that the link represents a match. Unlinked census records will be assigned probability of zero (0). By assigning probabilities, rather than removing matched records to resolve duplication, it gives the flexibility during data processing to either resolve or to aggregate the probabilities to adjust estimation.

It is anticipated that the duplication of records in more than one location would be most common in the cases of children with parents who live apart being recorded at both addresses.

The Census to Census Linking process will produce an output dataset that will not remove any duplicate records, but instead will contain assigned probabilities to be used in the Estimation and Adjustment process to make adjustments to the population count.

The Census to Census Linking assigns probabilities to linked records, and, hence, does not include a decision making as part of the process. However, the outputs of probabilities will be quality assured to ensure the process is running correctly.

In addition, the overall method described in the methodology paper[39] includes a contingency method for the case that administrative data is not available in 2022. This method would use probabilities calculated for each category of link from the 2011 census and a previous administrative data. In this circumstance, the same quality assurance checks will be applied using a 2011 dataset.

Method used in 2011

In 2011 only links where the name and date of birth agreed exactly were considered as potential duplicates. The method did not include checking against administrative data. The available administrative data was applied to a sample of census records, ensuring the sample contained enough duplicates to give an acceptably low coefficient of variation. Running the process on the whole dataset in 2011 would have been too computationally intensive.

Further adjustment of the population overcount used the Census to CCS[40] linking as part of the Estimation and Adjustment process. The records were assigned propensities within each stratum to be used as the in-census count for that record within Dual System Estimation (DSE). DSE is a statistical process using the links between two independent datasets of the same population to estimate the total population[41].

Additional matching of all linked census records to administrative dataset is a new methodology for 2022.

[39] For detailed methodology see: PMP015: Census to census linking and overcount correction | Scotland's Census (scotlandscensus.gov.uk)
[40] The Census Coverage Survey (CCS) is a voluntary, independent, post-enumeration, representative, sample survey used during coverage adjustment to produce population estimates.
[41] See the Estimation and Adjustment Methodology paper for more information on DSE: PMP001: Estimation and adjustment methodology | Scotland's Census (scotlandscensus.gov.uk)

Methods used by ONS and NISRA

Similarly, ONS proposes a number of approaches to resolve multiple records and resulting population overcount.

To estimate the expected scale of the error from multiple responses, ONS will use evidence from the linkage of the 2011 census data with the Longitudinal Study, together with the assumptions on changes in response patterns as a result of a greater number of online or individual responses.

5.9.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Sample check**

   A sample of linked records will be manually checked to ensure that the linking was performed correctly by the Census to Census Linking process. The sample will be assessed on the plausibility of the records being linked correctly according to pre-determined rules of the process as described in the methodology.

   It is anticipated that the process will link around 22,000 records. The manual quality assurance check will include a sample of around 100 links (~0.5%).

2. **Distribution of probabilities**

   In the Census to Census Linking process, linked records are assigned probabilities that are based on the scoring of the strength of those links on name and date of birth. Reviewing the distributions of these probabilities will reveal any potential unusual patterns and also will provide an indication of the overall level of duplication.

It is expected that the duplication of individual records in more than one location would be most common in cases where children whose parents live apart are being recorded at both addresses. However, any significant or unexpected deviations from these cases, for example, among records for adults, will require further investigation. This will determine whether these distributions of probabilities are due to errors in the process or represent a true composition of the data.

The nature of this process and statistical assessment does not allow confirmation of thresholds or expected shapes of distributions to be prepared in advance. Most of the decisions following this assessment will be completed during the processing of the 2022 census data. However, the distributions will also be compared to the equivalent 2011 data for initial reference.

In addition, the probability distributions will also be assessed for different subgroups, including children, other age groups, and by sex. Similarly, the distributions will be compared to the equivalent 2011 data for initial reference.

3. Checking the calculation

A further quality assurance check will be applied to ensure the main algorithm used in the Census to Census Linking is running correctly during the processing. To reduce bias, this check will be performed in a different environment to the main process using different statistical tool. Thus, the check will involve creating an Excel spreadsheet that will calculate the probabilities for the linked records independently to the main algorithm from the original dataset, which will be run using statistical software. The results then will be compared to the output from the main process to ensure there are no discrepancies.

These calculations as part of quality assurance will be tested on a sample. The sampling will be based on the grouping approach used in the main process. Grouping is used in the Census to Census Linking process for the initial linking to manage the large volume of records as the process is run on the full census

dataset. The grouping is not done on geographical location because duplicates can be recorded at different locations. Instead, the grouping is done on other main linking components – name and date of birth. Hence, the manual calculations will be applied to selected bands used in grouping rather than to the full dataset.

The details of the probability calculations used in the Census to Census Linking and the quality assurance process, are included in the main Census to Census Linking methodology[42].

5.9.3 Strengths and limitations

One of the strengths of the quality assurance for this process is including an additional calculation of the links probabilities outwith the main process. To account for practical implementation, it will be done on a sample. This check will provide a clear and independent assessment of the main algorithm.

However, one of the challenges of the process is that it is difficult to anticipate the expected distributions of probabilities that will confirm the correct running of the process. The approach of decision-making based on the data during the processing of the census data puts an additional pressure on providing prompt and acceptable resolutions. Therefore, the additional statistical support that will be provided by Census statisticians is invaluable at this stage.

This issue can be mitigated by preparing and confirming in advance the distributions of probabilities from 2011 census data and ensuring the statistical analysts applying the method are familiar with the process to assist with any decision making.

[42] For detailed methodology see: PMP015: Census to census linking and overcount correction | Scotland's Census (scotlandscensus.gov.uk)

5.9.4 Section summary

Census to Census Linking process identifies duplication in records where individuals are recorded usually living at multiple addresses, or at a wrong address. The process contributes to resolving the issue of population overcount in the census data.

To quality assure the Census to Census Linking process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.

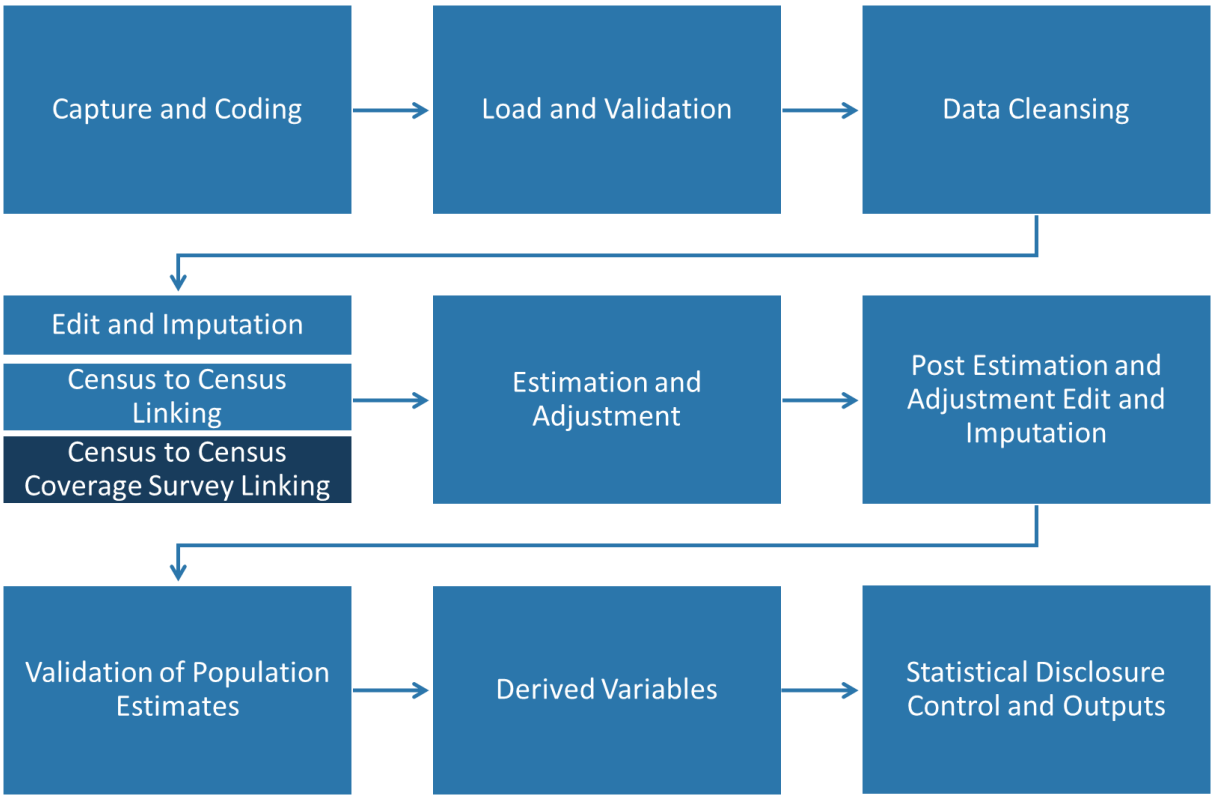
These include a manual sample check to ensure the linking process produces links that adhere to the methodology. A review of the distributions of probabilities of the links to reveal any unusual data patterns due to the potential errors in the process. In addition, a parallel calculation of the probabilities on a sample of the data will be compared with the main process algorithm to ensure they produce the same results.

5.10 Census Coverage Survey

To account for people and households who may not have been counted by the census, statistical methodology is used to estimate and adjust the coverage. This methodology is used to identify the number of people and households affected and to adjust the 2022 census estimates accordingly. The Census Coverage Survey (CCS) is a vital part of this process.

The CCS for 2022 will take place six weeks after the Census Day. This independent voluntary doorstep survey will aim to survey around 53,000 households (114,000 individuals) from around 1.5% of postcodes across Scotland. The CCS collects information about people and households that will then be matched to census records and used to estimate the size of the census undercount or overcount as part of the Estimation and Adjustment process (see Section 5.11).

Figure 16. Census Coverage Survey within the census data journey



The CCS questionnaire is different from the census and mainly collects the demographic data, such as age and sex. These data are used in combination with the census data to provide population estimates.

However, The CCS data will follow the same data journey as census data from collection to data cleansing. The CCS data will then be matched to the census data during the Estimation and Adjustment process. Hence, the intention is to use the same quality assurance methodology as for main census data processes to inform the quality assurance process for CCS.

The processes in the data journey for CCS are:

- Capture and Coding
- Data Cleansing (Remove False Persons, Resolve Multiple Responses, Filter Rules)
- Edit and Imputation
- Estimation and Adjustment (during this step the data is merged into the main census data journey)

For full details of the CCS methods, see the following methodology papers:

- Information about the Census Coverage Survey from the Scotland's Census 2022 website: Census Coverage Survey | Scotland's Census
- EMAP paper about how the Census Coverage Survey sample will be drawn: Scotland's Census 2022 - PMP002 - Census Coverage Survey (CCS) Sample Methodology (pdf)
- EMAP paper about the stratification during the Census Coverage Survey: Scotland's Census 2022 - PMP003 – Census Coverage Survey (CCS) – Sample Allocation and Reserve Sample Methodology (pdf)
- EMAP paper about how the sample of Communal Establishments will be drawn: Scotland's Census 2022 - PMP008 - Census Coverage Survey (CCS) - Communal Establishment Sample Methodology.pdf
- EMAP paper concerning how the Census data will be linked to the Census Coverage Survey data: PMP010 - Census to CCS linking - EMAP (scotlandscensus.gov.uk)

5.11 Estimation and Adjustment

Scotland's Census 2022 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 resulting in the undercount of the population. In addition, some people may get counted in the wrong place or more than once. The Estimation and Adjustment process aims to create population estimates fully adjusted for any undercount or overcount at both household and individual level for all geographical areas.

The Estimation element of the process estimates the level of undercount in population. It uses the results of the Census Coverage Survey (CCS). The Adjustment element of the process then adjusts this dataset to correct for the level of undercount. Adjustment generates additional household and person records and adds them to census returns to form a complete census dataset with counts that reflect the estimates produced. This complete dataset is further processed using post-adjustment edit and imputation process (see Section 5.8), and is used to produce statistical outputs.

The quality assurance checks for the Estimation and Adjustment process include:

- assessing the count of population in census and comparing it against the estimates from Census Coverage Survey (CCS);
- assessing demographic characteristics of the data and any unusual patterns to identify whether any potential differences between census and CCS can be explained or further investigated;
- using comparator data to quality assure estimates after each iteration.

5.11.1 Background and introduction

The Estimation and Adjustment process aims to create population estimates fully adjusted for any undercount or overcount at both household and individual level at all geographies.

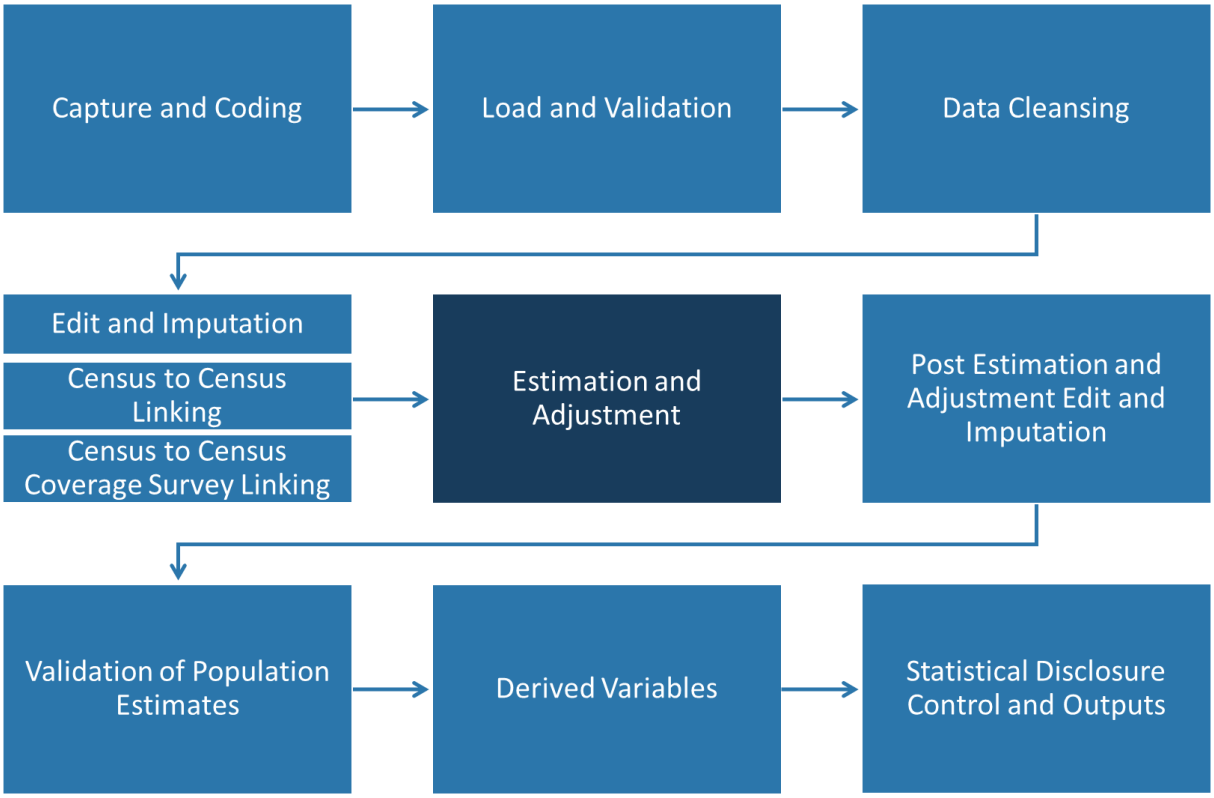
The Census Coverage Survey (CCS)[43], which is a voluntary, independent, post-enumeration, representative, sample survey, is an important part of this process. The Estimation and Adjustment process uses Dual System Estimation (DSE) [44] and the data from CCS to estimate the level of undercount. The primary 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. DSE is a statistical method that uses data from two independent sources: the census and the CCS. Thus, the DSE method requires statistical independence between the census and CCS, which is achieved by CCS using a different design and methodology from the census. The DSE process estimates the number of individuals and households missed in the census. 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.

For the Estimation and Adjustment process, the census data will be split into groups consisting of a few council areas. These are grouped based on similarity of demographics and their expected response rate. These processing groups, known as Estimation Areas (EAs), are used in order to create more manageable datasets with relatively consistent homogeneous levels of response.

[43] For more details on CCS, see Section 5.10.

[44] 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. See the Estimation and Adjustment Methodology paper for more information on DSE: PMP001: Estimation and adjustment methodology | Scotland's Census (scotlandscensus.gov.uk)

Figure 17. Estimation and Adjustment process within the census data journey



The Estimation element of the process applies the DSE to estimate level of census data undercount in CCS areas, and uses small area modelling to derive council area totals. This estimates how many households and people have been missed (including people in communal establishments).

The population count is estimated using the main estimation variables, a subset of the full census record variables. These include the data on age, sex, ethnicity, activity last week, and household tenure.

The Adjustment element of the process then uses the population estimated produced at the Estimation stage to adjust the census dataset to take into account of the undercount or overcount of population. This is done by creating synthetic households (or 'skeleton' records) to allow for those who have been missed by the census. The characteristics of these missed households are then imputed at the post-adjustment Edit and Imputation process using the census imputation system CANCEIS (see Section 5.8).

In addition to undercount, there will be cases of overcount where people have been counted more than once or counted in the wrong place. The process deals with this overcount by applying a down weighting factor to the population estimates. Furthermore, a national adjustment can be used to correct for an issue of the demographic spread of the population estimates[45].

The dataset created at the end of the Estimation and Adjustment process will represent the best estimate of the entire population. This adjusted dataset will be used to generate all statistical outputs from the census.

Method used in 2011

In 2011 there were a number of improvements made compared to Census 2001. These improvements included the measurement of overcount, adjustments for bias in the Dual System Estimation (DSE) estimates and increased use of external data, particularly as part of the quality assurance process.

In 2011, CCS did not collect data from communal establishments (CEs) with 100 or more bed spaces. Hence, administrative datasets were used instead. The coverage for these large CEs (such as prisons and university halls of residence) was measured using collected census data and administrative data[46].

Methods used by ONS and NISRA

ONS and NISRA are similarly planning to use Census Coverage Survey during the estimation process to match the number of people not included in the census. For

[45] For detailed methodology, see: PMP018: National Adjustment, Peer review and governance | Scotland's Census (scotlandscensus.gov.uk)
[46] For more details on 2011 methodology, see: 2011 Census Estimation and Adjustment Strategy | Scotland's Census (scotlandscensus.gov.uk)
For more details, see the coverage estimate and adjustment section here: 2011 census: Methodology | Scotland's Census (scotlandscensus.gov.uk)

more details on the methods used by ONS and NISRA see the respective 2021 Census quality assurance strategies.[47]

5.11.2 **Proposed method for quality assurance in 2022**

The Estimation and Adjustment process will determine whether the number of persons and the number of households reflect the expected number of persons and households in each Council Area. The quality assurance for this process will be applied to the two separate stages of the process: estimation and adjustment. The process includes the data input tables from both, census data and CCS data.

The following quality assurance checks will be applied during the processing of the data for Scotland's Census 2022.

Estimation

1. **Checking inputs**

   The pre-Estimation input tables will be checked by Council Area within census and CCS areas. Specifically, the number of households and number of persons in CCS postcodes and census postcode should be similar.

   The inputs will also be checked by examining summary statistics on a number of specific variables that can reveal unusual patterns and counts. These cases could include, for example, empty households, large only visitors households, only student residents in tem-time addresses, and communal establishments that are out of scope. All cases raising a concern will be further investigated.

[47] ONS: Approach and processes for assuring the quality of the 2021 Census data; NISRA: 2021 Census Quality Assurance Strategy

2. Summary statistics

CCS and census postcode clusters will be compared against the persons count and a decision will be made on which postcodes should be omitted. Clusters are the groups of postcodes, and are smaller than the estimation areas. The postcodes are omitted only in cases where there is an anomaly within a postcode, which will be removed for the estimation purposes. This is done, so that the process does not include these anomalies.

Pre-set CCS postcodes and clusters will be checked against the person count within both census and the CCS to make sure that the ratio between the two is close to one. This value indicates that the number of individuals and households is representative.

3. Checking process iterations

The number of people in each category of estimation variables (age, sex, ethnicity, activity last week, and household tenure) will be checked to ensure that when collapsing the variables the number of people in each category is not too small. This is done to ensure that variables within each category have a reasonable number to produce a smooth distribution. Previous methodology indicated that the count should not be fewer than five persons for each group. Therefore, quality assurance checks will involve making sure that the number of individuals for each category is in line with the data processing report produced for this quality assurance check.

Ratios by sex and age groups will be checked to ensure that any unusual patters can be explained or further investigated.

The population counts by clusters will be checked against DSE counts. A report will be produced outlining differences between the two sources and if and how the Estimation and Adjustment process dealt with these.

4. **Comparator data check**

Comparator data will be used to quality assure population count estimates after each iteration of the estimation process. This will be applied by estimation areas (groups of council areas). Demographic distributions by age and sex will be compared against the comparator source of NRS mid-year population estimates, and any large differences will be investigated.

The comparator data source is independent of the 2022 Census, and will be in line with the approach of Validation of Population Estimates process (see Section 5.12).

Adjustment

1. **Summary statistics**

Counts of imputed people and households will be checked for each council area to ensure that any noticeable differences in the data can be explained or investigated further. Counts of households within a postcode imputed in existing households (based on placeholders) will be checked against 'new' households that were created as part of the Adjustment process. The preference is to impute households into existing placeholders, so this count should be higher compared to the count of newly created households. Estimated figures before the adjustment stage of the Estimation and Adjustment process will be compared to the final adjusted figures to identify any unusual patterns in the data. Each quality assurance check will be captured in the process report.

2. **Process calibration statistics**

The Adjustment process uses probabilities of likelihood of people being missed in the census. These are then compared against the estimates. The weights will be adjusted so that the total weight for a category (individual and household) gets as close as possible to the corresponding estimate for that category. The probability

weights should match the estimates. Any instances where a mismatch is identified will be further investigated.

Further checks will be carried out to ensure that a particular person or a household is not used too many times as a donor for the new 'skeleton' records.

3. **Comparator data check**

Similarly to the same check at the Estimation stage, a comparator data of NRS mid-year population estimates will be used to quality assure the adjusted estimates counts and distributions by estimation areas and at Scotland level. This comparison will be used as a sense check to ensure no large discrepancies were introduced following the Adjustment process. Any discrepancies will be highlighted and investigated further.

5.11.3 Strengths and limitations

The process of Estimation and Adjustment is based on a very robust process and includes a number of automated processed with clear outputs that are straight forward to assess.

However, the approach of decision-making based quality assurance checks, in particular reviewing the omitted postcodes and comparator data check, during the live operations puts an additional pressure on providing prompt and acceptable resolutions. Therefore, the additional statistical support that will be provided by Census statisticians is invaluable at this stage.

This particular issue can be mitigated by preparing and confirming in advance the expected number of people and households for each category from 2011 census data or any other comparator sources and ensuring the statistical analysts applying the method are familiar with the process to assist the decision making.

5.11.4 **Section summary**

To quality assure the Estimation and Adjustment process during live operations for Scotland's Census 2022, a number of statistical quality assurance checks will be performed on the data.

These predominantly include reviewing the summary statistics produced during the process and comparing to the expected requirements of the pre-determined methodology. Quality assurance also includes reviewing the distributions of people and households, and comparing those to existing population comparator data to reveal any unusual data patterns that might occur as the result of the estimation process.

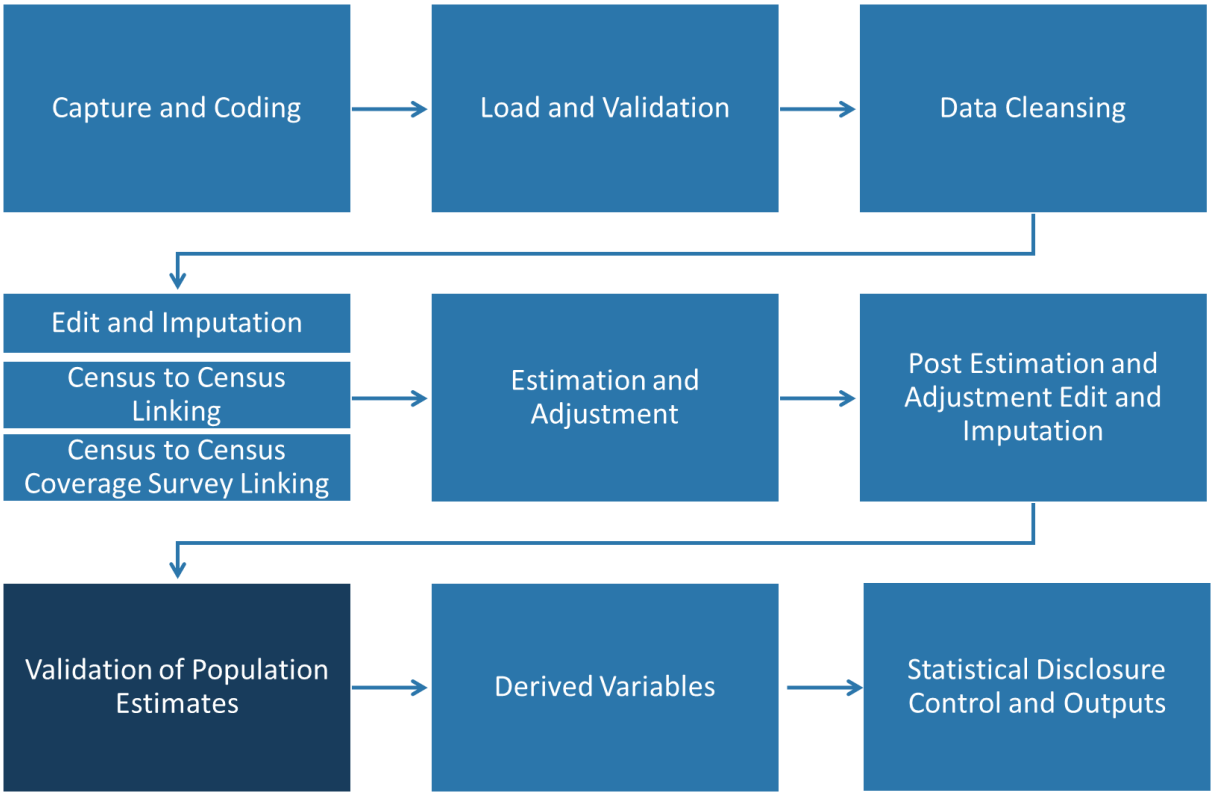
Some of these quality assurance checks are resource intensive. However, this can be mitigated by further development of the detailed methods in advance.

5.12 Validation of Population Estimates

Scotland's Census 2022 will produce population data for a diverse range of groups and characteristics, covering both individuals and households. Statistical quality assurance will be carried out throughout the processing of the census data. Validation of Population Estimates (VoPE) will be part of this process. VoPE ensures that the main demographic statistics are plausible for Scotland and its constituent Council Areas as well as other geographies before publishing the census data for public use.

VoPE process will compare census data with existing data and identify any instances where census data differ noticeably from these. The process will consider the likely accuracy of each comparator data source, including differences in collection, quality and known bias, and will focus on validating the census population estimates for National and Local Authority population counts.

Figure 18. Validation of Population Estimates within the census data journey



A suite of tools and methods will be used in order to focus on geographic areas, population groups and topic areas where there are inconsistencies or a need for further analysis. VoPE will use a range of comparator data sources such as:

- 2011 census data;
- administrative data;
- survey data or other sources.

Ongoing conversations with data suppliers and information from any official statistics sources will help with the understanding of the quality of these datasets. Knowing the strengths and limitations of these data will help when making decisions when conducting the VoPE checks.

The process aims to verify the census data given the comparator data. Characteristics of the census data to be checked and the comparator data will guide the type of check undertaken. There are three types of checks proposed:

1. **Range check** – check if the estimate lies within an acceptable range of the comparator data. This will take into account the quality and coverage of the comparator data.

2. **Proportion check** – check if the distribution and/or proportions present in the census data match with those in the comparator data. This is suitable for the comparator data that is based on a sample, and do not provide full coverage at national level.

3. **Ad-hoc comparison** – for some topics there is a lack of available comparator data sources. In some cases, the difference in methodological approach in collection of comparator data will not allow for direct comparison. For these topics, VoPE process will create bespoke methods of comparison based on coverage and quality of the available data.

At this point in the quality assurance process of the Scotland's Census 2022, census data are expected to be high quality, therefore, VoPE is a final validation check. Where discrepancies do arise there will be verification that the difference is adequately explained and if not, this will be investigated further.

The evidence to support population estimates will be reviewed by internal and external quality assurance panels of topic experts, who will advise on whether estimates are fit for purpose, require further adjustment, and/or help explain any anomalies identified.

For full details of the VoPE methods, see the following methodology paper: Statistical Quality Assurance - Validation of Population Estimates methodology paper | Scotland's Census (scotlandscensus.gov.uk).

5.13 Derived Variables

Following the processing of the census data, a number of output variables are created to undergo statistical disclosure control before being published for access by data users.

Some of these outputs are known as primary variables, which means the variables that are created directly from the responses to the census questions. In other words, primary variables are the variables that transform the census questions into electronic data. For example, the question on Number of Bedrooms will produce an output of a primary variable to show the data on number of bedrooms.

Derived variables, on the other hand, are created to supply census data users with additional variables that are not directly included in the standard set of questions. These additional variables are derived from the input variables from the responses to the census questions by grouping, combining, extracting, or calculating them. For example, the data on household size is created by calculating the count from the question on the number of people living in the household.

The quality assurance checks for the Derived Variables process include:

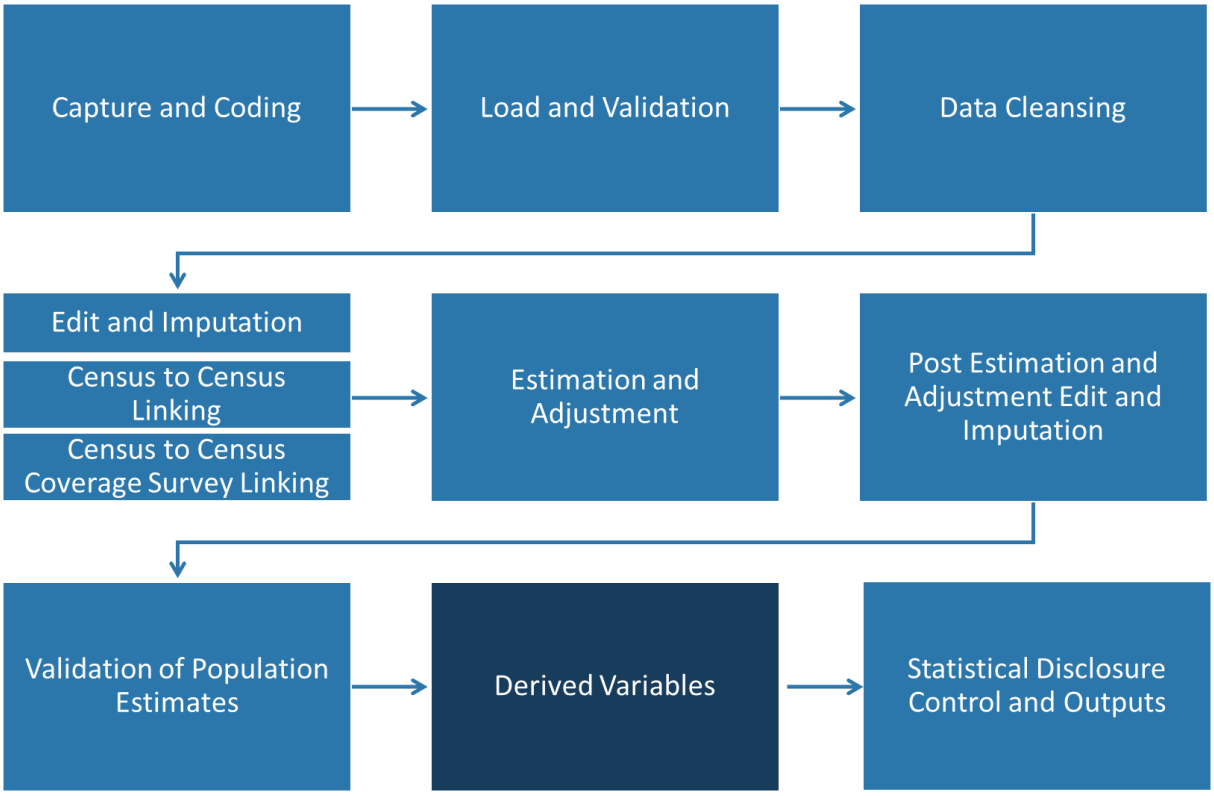
- review of summary reports of the process;
- manual sample check to ensure the correct logic was used to derive these variables;
- examining the distributions of the derived variables after the record swapping process.

5.13.1 Background and introduction

A derived variable is a variable that is computed from one or more variables. These will be incorporated into the census data journey and disseminated as part of the standard census data outputs tables and the census data flexible table builder. This will allow the stakeholders and data users to extract the census data directly from the Scotland's Census website.

Derived variables are created at the Outputs stage of the census data process and, as with all outputs, are subject to the Statistical Disclosure Control (see Section 5.14) to ensure the confidentiality of personal data.

Figure 19. Derived Variables process within the census data journey



Derived variables are created using the following methods:

- grouping responses into high-level groups;
- combining different elements of questions and/or other variables;
- extracting certain elements of questions and/or variables;
- calculating a count/number/date/year/distance from another variables.

An example for how a derived variable may be created, is grouping different, more specific professions into high-level groups. For example, creating an output on occupation from the question about person's job title; if the respondent answered 'physiotherapist' this will be grouped as 'medical professional' in the derived variable for occupation.

Information about each derived variable will be documented in a specification. This specification will be used as an internal audit trail and will include details on metadata, version control, input information, classifications, and logic.

Approximately 130 derived variables will be created during Scotland's Census 2022. The derived variables will be sorted into different tiers to indicate the status of their input variables. If a derived variable only consists of variables directly found in the questionnaire, this is a tier 1 derived variable as all the input variables are non-derived (or primary variables). A tier 2 derived variable includes tier 1 derived variables rather than just non-derived variables, and so on. In total, there will be five different tiers.

Method used in 2011

In 2011, the same underlying methods were used to create derived variables suitable for census data outputs. Following further stakeholder engagement into data user needs and the inclusion of new census questions, a number of new derived variables will be created for 2022.

Methods proposed by ONS and NISRA

ONS and NISRA will also be recreating some of the same derived variables that were used in 2011. In addition to this, both census offices are engaging with stakeholders to inform any new derived variables that may be useful to users. ONS, NISRA, and NRS are using similar process for creating the derived variables.

5.13.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be performed during the processing of the data for Scotland's Census 2022.

1. **Distributions report**

   The distributions of derived variables will be compared against 2011 data to investigate if the logic used to create the variables worked as intended. Some differences are expected, but if the distributions are considerably different, this will be investigated. This check also examines whether any expected derived variables are missing, or if any that were not planned have been added, as this may indicate an issue with the logic.

   If there are missing values, this may suggest a fault in applying the logic to the data. There are also a number of derived variables that have been newly created for 2022 due new questions in the census. Therefore, there are no 2011 census data comparison for these variables. To quality assure the derived variables for the new questions, alternative comparator data sources will be used. In addition, where possible, some comparisons will be made using 2021 census data from ONS and NISRA.

2. **Sample check**

   During the process of creating the derived variables, a manual sample check will be applied to assure that the logic works as intended. The derived variables will be compared to their input variables. These manual checks will be completed by tiers and will be applied to all the derived variables.

3. **Distribution check after record swapping**

   Similarly to all the census data outputs, derived variables will undergo statistical disclosure control to protect personal information in the census data. This

includes record swapping[48], which involves swapping the geographical information of a proportion of households that have unique identifiable characteristics. However, the record swapping is done in such way that the main structure of the data remains the same. This quality assurance check will compare the distributions from before and after record swapping process to ensure that it has worked correctly. This check will also ensure that no unlikely scenarios have been included. This is especially important for geographical questions (including Distance Travelled and Method of Travel) as distances between home and work may become unfeasible as a result of record swapping.

The overall report on distribution analysis and any required corrections will be peer reviewed by a group of Census statisticians.

5.13.3 Strengths and limitations

The process of creating derived variable is based on application of logic, which is a relatively straightforward process for quality assurance, as it does not require additional statistical analysis. However, the process will be resource intensive due to high number of the derived variables created.

One of the limitations of assuring the processes by comparing the data to comparator data sources is that there will not be exact comparison available for some of the variables. This will be the case for some of the new derived variables introduced in 2022 and when comparing with other countries' results. However, this can be resolved by conducting multiple comparisons for the variables where there is no single direct comparator source available.

[48] For more details on record swapping as a method of Statistical Disclosure Control (SDC), see Section 5.14.

5.13.4 Section summary

The Derived Variables process creates new variables by grouping, combining, extracting, and calculating the primary variables that are readily available from census questions. This is done to supply stakeholders and data users with as useful a dataset as possible.

To quality assure the application of the Derived Variables process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.

These include investigating the distributions of derived variables, checking for any missing variables, manually checking that the logic of all derived variables works as intended, and checking the distribution of particular geography-based questions after the record swapping process.

5.14 Statistical Disclosure Control and Outputs

National Records of Scotland (NRS) has a duty to ensure that the privacy of individuals and households is protected in all releases of census data. Statistical Disclosure Control (SDC) is used to protect personal information in the census data. These methods include making small changes to the data, controlling access to data, and controlling the level of detail that is available to the census data users.

Making small changes to the data provides a sufficient level of doubt on whether the values are exact or are the result of changes due to disclosure control. Controlling access to data ensures that the number of people who can access detailed personal data is kept to a minimum. All NRS staff who will have access to personal census data are subject to rigorous security clearance checks and are trained in statistical data management. SDC also controls the level of detail that is available to census data users. This way sensitive data that might lead to identification of individuals or households cannot be accessed.

There will be three main methods of SDC used for Scotland's Census 2022 data. This includes record swapping, cell key perturbation and flexible builder table rules.

The quality assurance checks for the SDC and Outputs process include:

These include providing peer review of reports of the quality assurance processes, such as:

- summary statistics of the processes;
- data utility checks and doubt checks;
- data distribution comparisons of datasets before and after the application of SDC methods;
- intruder testing exercise.

5.14.1 Background and introduction

Statistical Disclosure Control (SDC) refers to methods applied to census data outputs to protect the privacy of personal information. This includes making small changes to data, controlling access to data, and controlling the level of detail that is available to census data users.

SDC is needed to prevent the release of confidential information about an individual or household. NRS has a duty to ensure that the privacy of individuals and households is protected in all census outputs.

For Scotland's Census 2022, the three main SDC methods are:

1. **Record Swapping**[49]

   Record swapping involves swapping the geographical information of a proportion of households. For example, household 1 in area A is swapped with household 2 in area B. In any published census outputs the information from household 1 will be in area B and household 2 in area A.

   Record swapping was the main SDC method used to protect the privacy of households and individuals in outputs from Scotland's Census 2011.

2. **Cell Key Perturbation**[50]

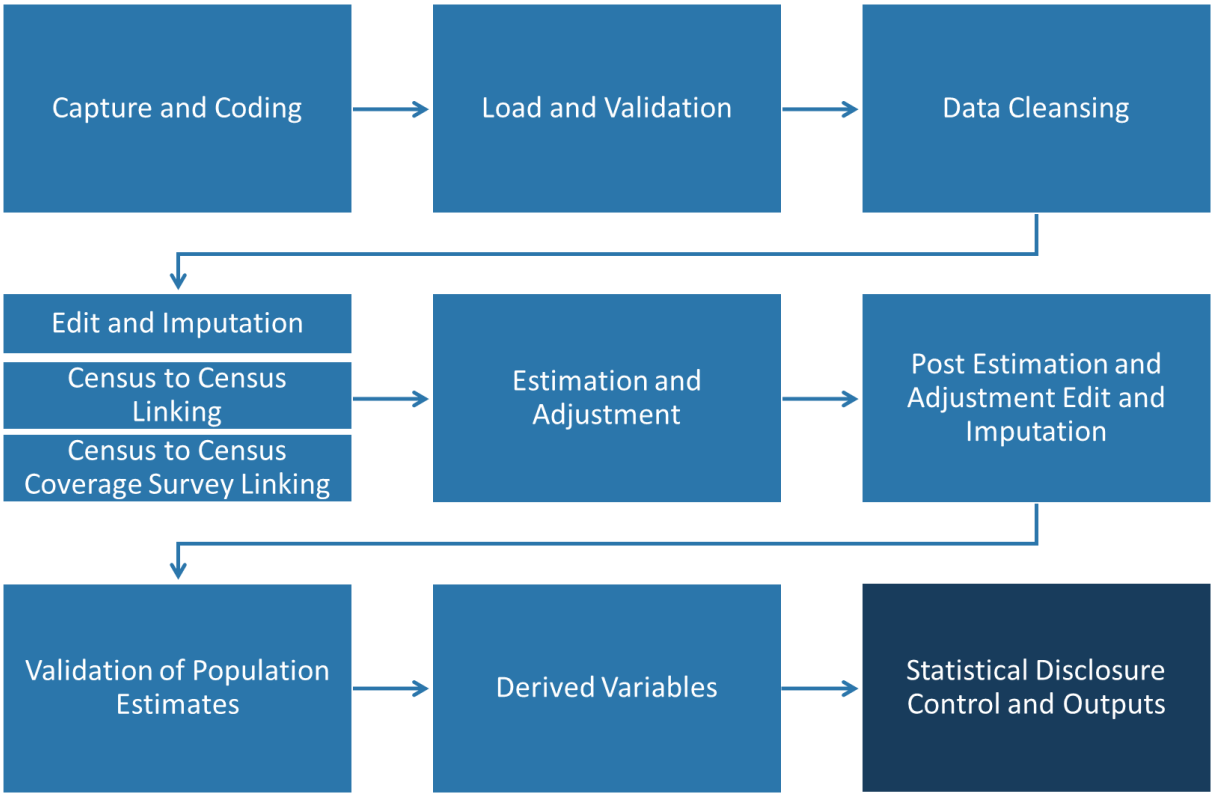
   A key innovation for Scotland's Census 2022 will be the availability of a flexible table builder tool. This will allow users to create their own tables from census data. In order to provide an additional layer of confidentiality protection, when data tables are created using the flexible table builder, a method of Cell Key Perturbation will be applied. Cell key perturbation introduces small adjustments to cells within output tables.

[49] For detailed methodology, see: PMP016: Household record swapping methodology | Scotland's Census (scotlandscensus.gov.uk)
[50] For detailed methodology, see: PMP017: Cell key perturbation | Scotland's Census (scotlandscensus.gov.uk)

3. **Flexible Table Builder Rules**

   The flexible table builder will also have built-in rules to ensure that tables containing information that might allow individuals or households to be identified cannot be accessed. For example, tables containing very small cell counts at a given geographic level will not be accessible in the table builder.

Figure 20. Statistical Disclosure Control and Outputs process within the census data journey



Method used in 2011

The main method of Statistical Disclosure Control (SDC) used in 2011 Scotland's Census was record swapping.[51]

[51] For more details, see the statistical disclosure control section here: 2011 census: Methodology | Scotland's Census (scotlandscensus.gov.uk)

Methods used by ONS and NISRA

Both ONS and NISRA are similarly proposing to use record swapping, cell key perturbation and table builder rules as methods for statistical disclose control for census data.

5.14.2 Proposed method for quality assurance in 2022

The following quality assurance checks will be applied during the processing of the data for Scotland's Census 2022.

1. **Record swapping**

The process applies to the whole Scotland census dataset, however, the dataset is split into batches of neighbouring Council Areas (CAs). Restricting the highest level of swapping to swaps between adjacent Council Areas ensures that households cannot be swapped over large distances to minimise the impact on data quality. For example, a household in the Scottish Borders could not be swapped with another in Orkney. The records are swapped in the Council Areas that are similar in terms of their location as well as urban and rural structure. This ensures the household characteristics and the data composition are correctly represented in the swapping process.[52]

The main task of the process is to introduce a level of uncertainty around cell values in the census datasets while minimising the impact on data quality or the underlying structure of the data. Hence, the comparisons of the pre-swapping and post-swapping datasets are built into the swapping process.

To quality assure the process, the datasets after the record swapping procedure will be assessed on whether swapping rates are above an agreed threshold at different geography levels (Output Area, Data Zone and Council Area). The exact thresholds

[52] For more detailed methodology, see: PMP016: Household record swapping methodology | Scotland's Census (scotlandscensus.gov.uk)

for swapping rates are not published and not widely known. This is to protect the SDC process from detailed examination of its construction to prevent any possible reverse engineering that might reveal any disclosive data.

However, these thresholds are pre-determined in advance and have been developed and approved in collaboration with the ONS and NISRA, as both use similar parameters. Specific rates also depend on the record swapping required for different areas. Hence, the swapping rate will not be the same across Scotland, but it will apply the same principle of SDC. This is applied in order to minimise changing of the underlying structure of the data. High volumes of record swapping can be detrimental to the original data composition, and thus, where possible, are not used.

Swapping utility measures will assess the level of data structure preservation against the volume of record swapping. Data utility measures assess the post-swapped data to ensure that the swapping process has not excessively changed the census dataset such that it is no longer useful for data users.

Similarly to the swapping rates, the utility metrics are highly sensitive and cannot be published nor shared widely within the Census programme.

The team implementing the SDC processes will perform a manual check on a sample of records at Output Area, Data Zone and Council Area geographical levels. This process will identify any changes to the demographic composition of different areas after the record swapping has been applied. This quality assurance check will also include an additional manual check of particular variables and households with very specific and unique characteristics. For example, data on minority ethnic groups.

A separate ‘doubt’ check will be carried out to ensure that the appropriate disclosure control has been applied. This check will use outputs tables created using swapped datasets. Doubt metrics give an indication into the level of certainty a data user or an intruder could have around whether or not a value in a cell in census output tables is real or has been altered during the swapping or imputation process. To protect data

confidentiality, this check will be performed exclusively by the team responsible for the SDC processes.

Due to the stringent and necessary confidential nature of the SDC process, other statistical teams within the Census Programme will not be able to access record swapping data directly. However, the high-level reports on the process and any required interventions will be made available for review and approval for overall statistical quality assurance. These reports will be produced at the Council Area (CA) and Scotland levels, thus avoiding disclosure of sensitive information, which is more likely at lower levels of geography.

Additionally, a review and evaluation of the data distributions before and after swapping process at Council Area and Scotland levels will ensure the correct application of the SDC processes. The swapping process is applied to protect records whose characteristics put them at risk of disclosure in census outputs. However, care is taken to preserve the main structure and composition of the original data. Hence, the distributions of the census data should not change throughout this process.

Further, an overall intruder testing will be applied to the output data following the record swapping and cell key perturbation processes (see the section below). This will test whether the SDC processes have sufficiently protected the data. The test will focus on high-risk variables, in particular protected characteristics, at the lowest levels of geography.

2. Cell key perturbation

Cell key perturbation is a method of statistical disclosure control that introduces small adjustments to cells within output tables. This is a new method for Scotland's Census 2022.[53]

Cell key perturbation is applied on a tabular level. Small changes are made to selected cells in the output tables. The main quality assurance for this process is performed by comparing perturbed and unperturbed tables, and checking that variables have been perturbed consistently across different tables. For example, different data users requesting the same data table will receive the tables with the same perturbation applied. This means that it will not be possible to isolate the perturbed figures by differencing the numbers from these tables.

Due to the nature of SDC, some of these checks will be carried out internally within the team responsible for the SDC process. Perturbed and unperturbed tables will be compared to ensure that perturbation is applied correctly and consistently to different variables. In addition, as part of this check, the same tables will be built multiple times to ensure that the results are the same each time they are created.

The perturbed output tables will also be cross-checked manually. A specially created processing environment will ensure that the statistical team performing this quality assurance will have access to outputs data only, and will not be able to see the perturbation table separately. This is to preserve the disclosure control of the data.

Further, statistical intruder testing will be performed to ensure the statistical disclosure process is sufficient. The tabular data can be output as Excel files to attempt to 'unpick' perturbed tables. There will be over 100 standard output tables; a sample of 10–15 will be sufficient to test whether the perturbation has been applied correctly. The 'unpicking' process itself can be automated by preparing the testing systems in advance, based on existing data.

[53] For more detailed methodology, see: PMP017: Cell key perturbation | Scotland's Census (scotlandscensus.gov.uk)

3. Flexible table builder and outputs

Following the required SDC procedures, the final census data outputs will be published on Scotland's Census website[54]. In addition to a number of standard tables being published, a flexible table builder will allow data users to create tables to their specific requirements.

There are a number of rules built into the flexible table builder to ensure that users cannot create disclosive tables. This includes limiting the total number of variables that can be included in a table, and restricting combinations of certain variables. Most of these rules are based on 2011 data, but additional rules were created for variables based on new census questions for 2022. The team responsible for the SCD process will test and develop these rules during the live Census operations, as the data for these new variables becomes available. It might be possible that the data for some of the new variables will require additional rules due to small population size. If required, any additional rules will be peer reviewed by statisticians within the Census Programme and relevant quality checks will be carried out to ensure that these rules have been applied correctly.

The accuracy of the standard data tables will be ensured by the standard quality assurance procedures employed for statistical publications. This will involve checks using statistical software to ensure that the published data tables are accurate and consistent with the census output datasets. These checks will be performed by internal statistical teams within the Census Programme.

5.14.3 Strengths and limitations

The nature of the Statistical Disclosure Control (SDC) process means that the scale of independent assessment from a different statistical team is strictly controlled. This is to protect the confidentiality of personal data, which is one of the main objectives of Scotland's Census. For this reason, the implementation of SDC procedures and

[54] For more information see: Home | Scotland's Census (scotlandscensus.gov.uk)

the detailed quality assurance of outputs following these methods, are restricted to a limited number of statistical analysts within the Census Programme.

Statistical analysts responsible for the SCD process perform a number of rigorous data quality assurance checks. This will be supported by a secondary statistical team within the Census Programme assessing the overall structure of the data to ensure the SDC methods have been applied as specified in the agreed methodology, and in addition to performing a separate intruder testing to ensure the SDC methods have provided a suitable level of data confidentiality.

5.14.4 Section summary

The application of Statistical Disclosure Control (SDC) to the census data is required to ensure the confidentiality of personal data for individuals and households. Certain details within the SDC methodology are not widely known in order to protect personal data and ensure that the methods are not reversed to disclose any sensitive information.

To quality assure the application of the SDC process during live operations for Scotland's Census 2022, a number of pre-determined statistical quality assurance checks will be performed on the data.These include providing peer review of reports of the quality assurance processes, such as:

- swapping rates for the record swapping process;
- data utility checks and doubt checks to ensure that the appropriate disclosure control is applied while ensuring the data structure and functionality to data users is preserved, and;
- data distribution comparisons of datasets before and after the application of SDC methods to ensure that the methods have not had an impact of the population counts and the overall structure and composition of the data.

Furthermore, an intruder testing exercise will ensure that the SDC methods are protecting the data, and that users cannot undo any of this protection.

6. Conclusion

Scotland's Census collects information from every person and household in Scotland, and produces essential data, which is not available in any other data sources. A range of stakeholders and data users use the census data at national and local levels for policy development, resource allocation, research and so on. The quality of the data outputs is vital and is expected to be of high quality.

The approach of Statistical Quality Assurance, and Assurance of Processes in particular, is centred around assessing that data processing for the census data is applied according to process methodologies, and ensuring data quality throughout.

This paper has provided the details of the quality assurance checks for each statistical process within the census data journey. The main strength in the overall Assurance of Processes approach is in offering additional level of assessment and quality assurance. This comes from combined efforts from the internal Census teams that are creating methodologies and applying processes during live operations, as well as peer review and support from a dedicated statistical quality assurance team, who are separate from the process operation.

This approach will provide an additional level of independent assessment and quality assurance, will expand the coverage of the data being checked and reduce the likelihood of potential biases affecting the quality assurance process.

7. Annex

7.1 Glossary

Term	Definition
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. Administrative data are collected with a specific decision-making purpose in mind pertaining to an individual, and so the identity of the unit corresponding to a given record is crucial. In contrast, in the case of statistical data, on which no action concerning an individual or a business is intended or even allowed, the identity of individuals/businesses is not of interest.
Administrative Data Sources	ONS 2011 Definition: Administrative data sources are collections of data maintained for a purpose other than the collection and production of statistics. These sources are typically managed by other government bodies. A range of data was used to help quality assure 2011 Census estimates. These data included administrative sources (such as the number of people registered with a doctor, the number of households registered for council tax purposes), surveys (such as the Integrated Household Survey) and other official ONS population products (such as the mid-year population estimates).

Assurance of Processes	Quality assurance activities at each step of the Census data journey.
CANCEIS	Canadian Census Edit and Imputation System: Software designed by Statistics Canada.
Capture	Capture is the process by which a return is converted into a suitable electronic format (in the case of paper returns) and is matched to an electronic template ready for coding. Capture is either paper capture (from paper returns) or online capture (from online returns).
Cell Key Perturbation	A statistical disclosure control method that involves adding or taking away a small number to or from selected cells of a table in a statistical way that does not introduce bias and is replicable.
Census	The official count of every person and household in the country on a given day.
Census Coverage Survey (CCS)	The Census Coverage Survey (CCS) is a voluntary, independent, post-enumeration, representative, sample survey used during coverage adjustment to produce population estimates.
Classifications	A statistical classification is a set of categories, which may be assigned to one or more variables and used in the production and dissemination of statistics.
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.

	There are three types of coding: • Point of Contact • Traditional • Derivation Coding can occur in two ways: • Automatic • Manual
Communal establishment (CE)	A communal establishment is typically a managed residential accommodation where there is full-time or part-time supervision of the accommodation. There are about 4,600 communal establishments in Scotland. Types of communal establishment For Scotland's Census communal establishments are grouped into four types. There will be a different approach to collecting census data for each type. • Type 1: care homes, staff accommodation and religions establishments • Type 2: hospitals, prisons, schools and children's homes, hotels, guest houses and hostels • Type 3: student halls and accommodation, defence establishments • Type 4: homeless people in temporary accommodation, rough sleepers For more details, including collection methodology, see: Contact with communal establishments | Scotland's Census (scotlandscensus.gov.uk)

Council Area (CA)	Councils (local authorities) form the single tier of local government in Scotland. There are 32 councils (local authorities) in Scotland, the administrative units of local government.
Data cleansing	A collection of processes applied to census data to account for specific errors, and prepare the data so it is suitable for later statistical processes.
Data quality	Data quality refers to the condition of a set of values of qualitative or quantitative variables. Quality or 'fitness for use' of statistical information is defined in terms of six constituent elements or dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence.
Data Zone	NRS Geography defines 2011 Data Zones as groups of Census Output Areas. Data zones have populations of between 500 and 1,000 household residents. To create 2011 Data Zones, Scotland was divided into 6,976 Data Zones, which are the main geography used for small area statistics. They were first created in 2004 by combining 2001 Census Output Areas, as a way to monitor and develop policy at small area level. Each Data Zone has approximately the same population (750), but because they are population based, they can vary greatly in size of area.
Derived Variable	A variable, which has been created using other variables.
Donor Imputation	Copying values from another 'donor' record into the failed record.
Donor record	A record which is used to help impute a failed record. Response values are copied from the donor to the failed

	record in order to replace missing or inconsistent responses, or resolve inconsistencies.
Edit and Imputation	The detection and repair of gaps or inconsistencies in census data, to ensure a complete and consistent dataset. Edit: the detection of missing, invalid or inconsistent responses. Imputation: the correction of missing, invalid or inconsistent responses.
Estimation Area (EA)	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.
Failed record	A record which contains missing, invalid, or inconsistent responses.
Hard edit	A rule which defines something which is impossible or so rare that most occurrences are errors. A hard edit specifies things which will not be allowed in the dataset.
Household	• One person living alone, or • A group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room or dining area. A household may also be:

	• a person or a group of people living in sheltered housing or very sheltered housing (irrespective of whether there are other communal facilities), • a person or a group of people living in a temporary or mobile structure (for example a caravan, mobile home or boat) on any type of site that is their usual place of residence.
Imputation rate	For a variable is the proportion of submitted returns where that variable has been imputed due to missing or invalid values, or inconsistencies. It should be made clear to external users whether deterministic changes are included in this count.
Inconsistent response	A response to a question, which contradicts other information given.
Invalid response	Where a response was provided, but it is not an acceptable value.
Manual coding	Manual coding is where a variable's value is assigned a code by an operator who follows a set of guidelines to determine the correct code to assign. Manual coding occurs when automatic coding fails to assign a category to a value.
Match	Two records that represent the same individual.
Metadata	Data that defines and/or describes other data.
Missing response	Where a respondent has not answered a question and a response was required.
Non-match	Two records that represent different individuals.

Northern Ireland Statistics and Research Agency (NISRA)	Northern Ireland's census is run by the Northern Ireland Statistics and Research Agency (NISRA).
National Records of Scotland (NRS)	National Records of Scotland (NRS) is responsible for planning and carrying out the census in Scotland.
Office for National Statistics (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.
Outputs	All numbers, tables, graphs, maps and text that show or describe the results of the census. This includes all supporting information and metadata.
Routing	Routing ensures that respondents only answer relevant questions, based on their answers to screening/previous questions. In automatic routing, respondents skip a question without seeing the question they are skipping.
Scanning	The process to create a set of digital images from a paper questionnaire.
Scotland level	As well as the other census geographies, statistics are available for Scotland as a whole; these are referred to as 'Scotland level'.
SIC	Standard Industrial Classification. Codes assigned to industry categories.
SOC	Standard Occupational Classification. Codes assigned to occupation categories.

Soft edit	A rule which defines something which can be considered to be an outlier. A soft edit specifies things which should not be disproportionately propagated throughout the dataset as a result of imputation.
Statistical Disclosure Control (SDC)	All methods applied to census outputs to protect the privacy of personal information. It includes making small changes to data, controlling access to data, and controlling the level of detail that is available to census data users.
Statistical Methods and Data Processing (SMDP)	The process of turning census response data into fit for purpose unit record data. It involves coding, editing, imputation, reconciliation, applying filter rules, estimation, adjustment and deriving variables.
Swapping rate	Proportion of households that are selected for swapping as part of statistical disclosure control. However, the swapping rate is not published to ensure swapping cannot be unpicked.
Record swapping	This is a method of statistical disclosure control that is used to protect the confidentiality of respondents data. It involves swapping the geographical locations of records. For example, household 1 in area A is swapped with household 2 in area B. In any published data the information from household 1 will be in area B.
Unsubmitted return	An unsubmitted online return is an online return where the respondent has not completed the online collection process by submitting their questionnaire responses to National Records of Scotland (NRS).