Scotland’s Census 2022

Statistical Methods and Data Processing

Rehearsal Evaluation

July 2020

Contents

1. Introduction

1.1 Scotland's Census 2022

The census is the only official count of every person and household in Scotland and the only questionnaire of its kind to ask everyone the same questions at the same time. For more than 200 years, the census has reflected the society in which we live, a snapshot in time of who we are as a nation, and how we live and work together. Crucially, the census still remains the best way to gather vital information that helps improve the lives of those living in Scotland. By providing this information, public, private and third sector organisations are able to plan and better deliver public services for all in Scotland.

It was announced on 17 July 2020 that the date of Scotland's next census would change from 21 March 2021 to 20 March 2022, due to the impact of COVID-19 on vital preparations for the census. This will be the 22nd census to take place since 1801 and the 17th to be managed independently here in Scotland.

Given we are now firmly into the digital century, as part of the digital first approach across the public sector, the next census will be the first one to be conducted primarily online. However, those that wish to complete a paper questionnaire will still be able to do so.

Either online or on paper, each household will be required to complete a household questionnaire, which contains questions about:

- the household as a whole ('household questions'), and
- each person usually resident in the household ('individual questions').

The householder will normally be responsible for providing the information on the questionnaire for the whole household. Any household member aged 16 or over can request an individual questionnaire in confidence. The individual can then maintain their privacy by providing their information in private.

People who live in 'communal establishments', such as hotels, hospitals and care homes, will complete individual questionnaires.

1.2 Scotland's Census Rehearsal 2019

Delivering a successful census for all is not a straightforward exercise. That is why, as part of our work to achieve this, National Records of Scotland carried out a major public rehearsal exercise that tested some of the systems and services that will be used in the next census.

The rehearsal data collection took place between 7 October and 7 November 2019, using a reference day of 13 October, and was conducted in three local authority areas; namely parts of Glasgow City, Dumfries and Galloway and Na h-Eileanan an lar. These areas were selected to allow National Records of Scotland to test approaches in rural, urban and diverse communities.

Participation in the rehearsal was on a voluntary basis. Over 72,000 households were contacted with nearly 18,500 responding. The rehearsal did not run a test of the Census Coverage Survey (which is a survey that takes place after the main census has been completed and is used to help measure the quality of the census), nor the count and listing of communal establishments (e.g. student halls of residence, care homes, hospitals etc.).

The rehearsal was an important milestone for Scotland's Census 2022, and its success has allowed us to gain vital insights on the operation of the systems and processes used to gather census information, helping us to discover what works and where improvements are required. This information was published in an evaluation report on 31st March 2020.

National Records of Scotland recognises this could not have been achieved without the goodwill of those that participated in the rehearsal, particularly members of the public and our external partners and stakeholders and we would like to thank them for their help with this important work.

1.3 Statistical Methods and Data Processing Rehearsal

The statistical methods used and how the data is processed, hereafter referred to as Data Processing, is what takes place after census data has been collected, captured and coded. Data processing for Scotland's Census 2022 is key to ensuring the data we collect is clean and complete to accurately represent Scotland's population. The statistical methodologies involved in this are broadly defined as Data Cleansing, Edit and Imputation and Estimation and Adjustment.

Where appropriate, administrative data is used to quality assure the data processing steps. This is done by linking the census data to administrative data after it has been through the processing step and before it is passed to the next stage. This acts as an independent check for potential inconsistences in the key variable date of birth and ensures true records are not removed from the dataset.

A simplified representation of the census data journey is shown in Figure 1. Data Processing covers all areas from Data Cleansing through to Output Preparation. Quality assurance checks using administrative data are built into various processes in the data cleansings steps and the census to Census Coverage Survey linking which takes place during the estimation step.

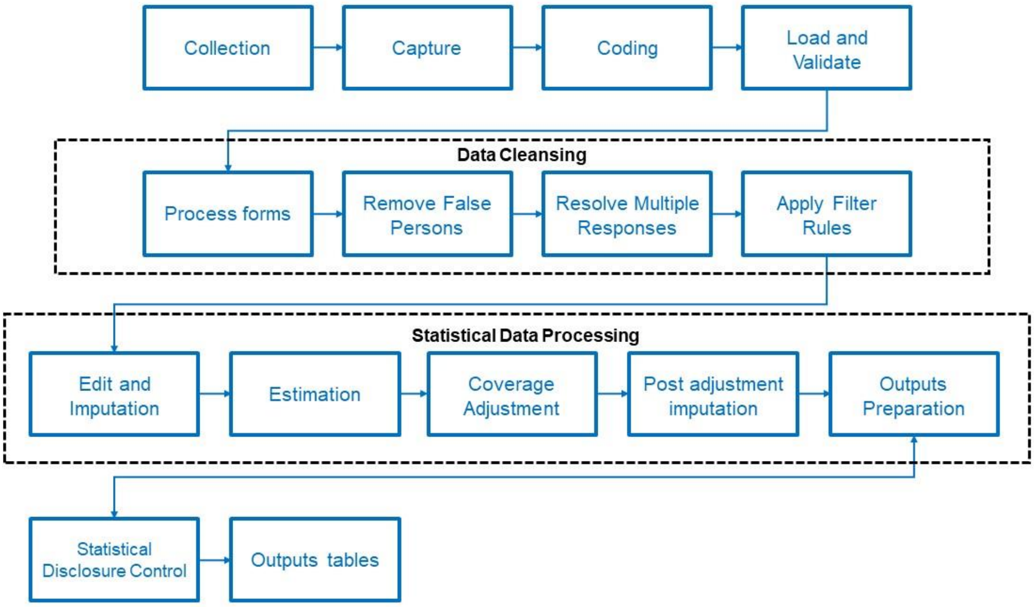


Figure 1: An overview of data processing in 2022

The data collected during Scotland's Census Rehearsal 2019 was used to test the Statistical Methods and Data Processing steps for Scotland's Census 2022.

The Statistical Methods and Data Processing Rehearsal was carried out from April 2020 to June 2020. Six data processing steps were tested. These were:

- Remove False Persons (RFP)
- Resolve Multiple Reponses (RMR)
- Filter Rules
- Name Re-ordering – on paper questionnaires only
- Edit and Imputation (E&I)
- Estimation and Adjustment (E&A)

Rehearsing the use of administrative data was also carried out alongside the six main steps of the Data Processing Rehearsal. The processes tested, which support the Statistical Methods and Data Processing steps, were:

- Remove False Persons (RFP) – to check the quality of this process
- Resolve Multiple Reponses (RMR) – to check the quality of this process
- Date of Birth Checks (Missing and Different) – to support Edit and Imputation
- Census to Census Coverage Survey (CCS) Linking – to help check for those who may have been missed in the census
- Census De-duplication – linking the census to itself to identify duplicate entries

The administrative dataset used for the rehearsal was NHS Central Register (NHSCR). This data was a snapshot taken as of 30 June 2019. The Data Protection Impact Assessment (DPIA) and the Quality Assurance of Administrative Data (QAAD) will be published alongside this paper.

This report provides a summary of the key findings of the Statistical Methods and Data Processing Rehearsal. It also outlines the next steps for National Records of Scotland to undertake to ensure that timely and reliable data processing will contribute to the successful delivery of Scotland's Census 2022 outputs.

2. Summary

The rehearsal showed that all the processes tested were able to run and complete successfully. In all of the processes tested it was found that using administrative data added value to the quality assurance process of the data and data processing steps.

Some changes to the software coding and to the format of the rehearsal data had to be made to allow the processes to be tested, however the rehearsal provided reassurance that the methodologies that have been developed so far worked well.

The rehearsal also allowed for a clerical review of the method for linking datasets, which is needed in each of the processes which use administrative data. These links are marked for automatic acceptance or for manual checking. For rehearsal, all links were reviewed (both automatic and manual). This thorough approach validated the linking methods being used and found that almost all the 'automatic' links were correct. As part of this clerical review, spreadsheets were developed in Excel which were fully tested and evaluated as part of this rehearsal.

The rehearsal also highlighted some areas that need further development or improvement. These include:

- Updates to the format of the data for live census to ensure it is suitable for data processing
- Further development of quality assurance and clerical review processes
- Further development of the sequencing of the data processes
- Further updates to the coding software and finalisation of methodologies

3. Key Findings

Ahead of the rehearsal, National Records of Scotland identified specific processing steps to evaluate. The following sections provide a summary of each processing step, what questions were addressed during the rehearsal, including the key findings and the further work required for each.

3.1 Remove False Persons (RFP)

Remove False Persons, or RFP, is a data cleansing process which looks at possible false person records in census data processing. This is done using two methods:

- The 2 of 6 Rule
- Name Check Rule

The 2 of 6 Rule was developed to detect and remove false records. This rule states that to be considered a true record, a person must have filled out at least two of the following six variables (one of the two variables must also be either of the name or date of birth variables):

1. Name in the individual questions
2. Name in the relationship matrix and household questions
3. Date of Birth
4. Relationship to others on household
5. Sex
6. Marital Status

In cases where a person has failed 2 of 6, but has included either a name or a date of birth, it may be possible to use administrative data records to check the plausibility of a person's existence, which allows us to retain these partial records as a true record.

The Name Check Rule will filter for obviously falsified names (such as 'Anonymous' or 'No one') or other indicators that there is not actually an individual on a record. Where a record has a falsified name, this will be turned into 'missing' and the record will then go through the 2 of 6 process.

During the rehearsal administrative data was used to help quality assure the RFP processing step. The 2 of 6 Rule was applied to the rehearsal data set to try to find cases which would have not successfully passed the RFP processing step, creating a simulated sample. This sample contained records which either had a valid name or date of birth. The intention was to take these records and attempt to link them to the administrative data source using the name or date of birth. This would have given us evidence the record was likely to be true and should be considered for inclusion in the data. Given the small rehearsal sample of around 40,000 cases and the filter rules for the RFP cases, there were none which met the criteria to run this process.

Cases passed for review would have been clerically reviewed using the clerical review sheet developed in Excel.

However, when this method was tested on the 2011 census data (with a sample of around 500,000) only 43 cases were identified after clerical review. This is entirely reasonable due to the lack of information in these records. Based on the work using the 2011 census data, when the numbers were scaled up to the size of a full census, it was estimated that approximately 430 records could be kept in the census dataset that would have otherwise been discounted. This additional quality assurance using administrative data would improve the quality of the census data, while having a small impact on resources and timings of processing runs.

Questions addressed during rehearsal	Key findings	Next steps
Is the data supplied in the correct format?	Data transformation was required to allow the data to be processed through RFP. Some workarounds had to be written into the RFP code to allow the process to run as expected.	Further work is being undertaken to ensure updates are made and the data from live census will be in the format required for data processing.
Does the RFP process complete successfully?	Yes. The process was able to run successfully and output results.	The rehearsal has highlighted parts of this process that still have to be developed prior to live census, such as how to carry out manual review of the data. This work is now in development.
Are the results of sufficient quality?	Respondent behaviour for the rehearsal is not indicative to that of a census, as the rehearsal is voluntary and the census is not. This means statistical quality is difficult to assess. However, the data was able to be processed successfully, with the data transformations mentioned above. Insufficient numbers of the most common 'false' name strings[1] were found (only a handful per forename/ surname combination), which was somewhat expected.	Further work will be needed to develop the 'false' name strings process before census live. Work on developing the quality assurance methodologies and a quality assurance dashboard for this process is also underway

[1] 'False' name strings are common ways for a respondent of a questionnaire to express information, but do not attribute themselves to a true record, e.g. Anonymous, No one, none.

Do the benefits of using administrative data outweigh costs?	Yes. There are clear benefits of using administrative data to quality assure this step. If done as a by-product of quality assurance on other processing steps, it allows for an increase in data quality while not having a big impact on resources and timings.	Further work on the sequencing (the order in which steps are run) with the data processing system, outputs and inputs is required to understand how the processes will run from end to end during live census.

3.2 Resolve Multiple Responses (RMR)

The Resolve Multiple Response (RMR) process is a data cleansing procedure that attempts to reconcile duplicate records of communal establishments, households and people which inevitably appear in the census dataset. For example, where a household may have inadvertently filled in the census online and on paper.

This process involves matching the census dataset to itself based on key variables (such as name and date of birth) to identify potential duplicate records. These records are then reviewed and are either reconciled into one record (if considered to be the same person) or are both kept (if considered to be different people).

Administrative data was used as an additional element of quality assurance on this process.

As noted previously, the basis of the RMR processes is that the census is linked to itself and looks for any duplicate records within a given postcode. Following this, the census was linked to another administrative data source to help identify duplicate records. The aim of this process is to resolve these duplicate records, reduce the potential overcount of the population and therefore the error in the census dataset. The work undertaken, as with the RFP process, included the development of Excel spreadsheets for both a clerical review and then a full review of all links found.

Questions addressed during rehearsal	Key findings	Next steps
Is the data supplied in the correct format?	Data transformation was required to allow the data to be processed through RMR. Some updates had to be made to the RMR code as well to allow it to process rehearsal data.	Further work is being undertaken to ensure the data from live census will be in the format required for data processing.

		Further review and update of the RMR code is required to bring it in line with updated coding specifications[2]. A process for accessing management information[3] in RMR has to be determined. Further work on how to link specific duplicate records together is also required. For example, where an address has more than one distinct household and an individual questionnaire is received for that address, a process to decide which household it should be joined to is needed.
Does the RMR process complete successfully?	Yes. The process was able to run successfully and output results, although it was limited by the rehearsal scope.	Further development is required in terms of decisions deferred until after rehearsal (e.g. resolution of within-postcode matches), items out of scope for rehearsal (e.g. communal establishments) and harmonisation with the Office for National Statistics (ONS) and Northern Ireland Statistics and Research Agency (NISRA) RMR methodology. This work is in progress.

[2] The coding specifications detail how to code questionnaire responses in a format suitable for analysing.

[3] Management information (MI) is used to measure and manage operational delivery during live census. An example of MI is live returns rates compared to forecast return rates.

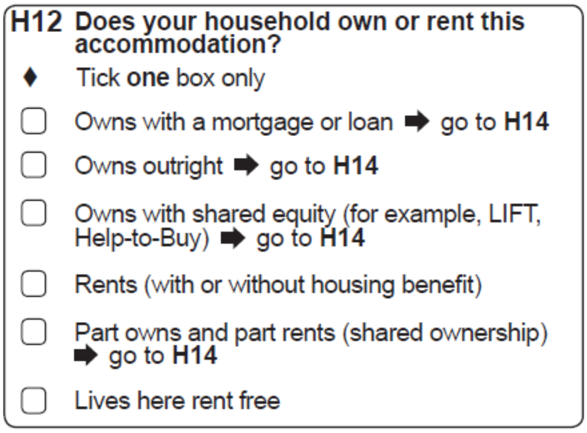
Are the results of sufficient quality?	Respondent behaviour for the rehearsal is not indicative to that of a census, as the rehearsal is voluntary and the census is not. This means statistical quality is difficult to assess. However, the data was able to be processed successfully, with the data transformations mentioned above. The rehearsal highlighted a gap in our understanding of how record identifiers should be assigned and retained.	Further work to develop suitable record identifiers and then make updates to the RMR process is required. Work on developing the quality assurance methodologies and a quality assurance dashboard for this process is also underway
Do the benefits of using administrative data outweigh costs?	Yes. The rehearsal showed more cases needing resolved than expected, approximately 1 per cent of the sample. This is partly due to the switch to online data collection. For example, where a respondent had already done a partial return online then requested a new password and submitted a separate full return. This process reduces duplication of records due to multiple returns from the same person.	For improved quality assurance for census 2022, the RMR process will be run with administrative data linking and clerical review. As with other processes, further work on the sequencing with the data processing system, outputs and inputs is required to understand how the processes will run from end to end during live census. The layout of the clerical review spreadsheets are being looked into to optimise their performance for the 2022 census.

3.3 Filter Rules

In census, Filter Rules is a process which resolves issues in the answering path (routing) of a questionnaire. In 2022, this will be required for paper questionnaires.

When a census paper questionnaire is received and scanned, each question is captured and coded separately. Filter Rules are used as part of the data cleansing process to resolve routing errors or issues in a record once coded answers are combined again. For example, not everyone is required to answer each question, in which case the questionnaire directs the respondent to the next appropriate question. However, sometimes the guidance is missed, or people simply wish to answer the question. Conversely, sometimes people do not answer questions they are directed to. When the data is coded, blank questions are automatically marked as “Missing”. The filter rules process checks to see which related questions have been answered and changes those which do not agree with the routing rules to “No Code Required”[4].

An example of this is below. The landlord and tenure filter rule resolves routing issues between the answers on question H12 (Does your household own or rent this accommodation?) and H13 (Who is your landlord?).



H12 Does your household own or rent this accommodation?

♦ Tick **one** box only

□ Owns with a mortgage or loan ➡ go to **H14**

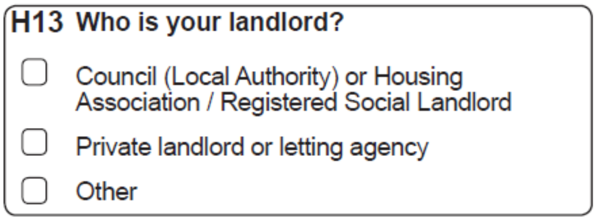
□ Owns outright ➡ go to **H14**

□ Owns with shared equity (for example, LIFT, Help-to-Buy) ➡ go to **H14**

□ Rents (with or without housing benefit)

□ Part owns and part rents (shared ownership) ➡ go to **H14**

□ Lives here rent free



H13 Who is your landlord?

□ Council (Local Authority) or Housing Association / Registered Social Landlord

□ Private landlord or letting agency

□ Other

If either, “Owns outright” or “Owns with a mortgage or loan” are ticked, question H13 should be skipped (for online respondents, question 13 would not be seen as people will be automatically routed past it). The rule resolves those who tick one of the two options but go on to answer *Who is your landlord?*

Questions addressed during rehearsal	Key findings	Next steps
Is the data supplied in the correct format?	Data transformation was required to allow the data to be processed through Filter Rules.	Further work is being undertaken to ensure updates are made and the data from live census

[4] All original responses provided by census respondents will be captured and archived. Any changes made to a record during the filter rules process will be fed through to the final outputs dataset.

	Unique identifiers for individuals and households had to be created because the software used in the analysis requires a numeric variable.	will be in the format required for data processing. Further work is required to establish how unique identifiers for individuals and households will be created in live census.
Does the Filter Rules process complete successfully?	Yes. The process was able to run successfully and output results.	Further updates to the Filter Rules code are required before live census. The rehearsal has highlighted parts of this process that still have to be developed, such as development of a flag strategy[5].
Are the results of sufficient quality?	Respondent behaviour for the rehearsal is not indicative to that of a census, as the rehearsal is voluntary and the census is not. This means statistical quality is difficult to assess. However, the data was able to be processed successfully, with the data transformations mentioned above.	Work on developing the quality assurance methodologies and a quality assurance dashboard for this process is underway.

3.4 Name Re-Ordering

This is a data cleansing process that was developed and tested during the rehearsal.

For a census return, respondents are asked to add people to the household questionnaire in the same order that people will appear in the individual questions

[5] A "flag strategy" is the approach that we take to highlighting records of concern, so that we may locate, review, and quality assure them – for example, if the records appear as they should, or if the correct changes were made. Flag strategy includes the criteria which we use to create them (specific rules are used to impose consistency) and the names of the variables we use for flags.

i.e. the first person listed as usually living in the household is also 'person 1' in the household and individual questions.

However, sometimes this does not happen. Part of the rehearsal was to develop a method that compared the names on the household questions to the names on the individual questions and then tried to make sure they were matched up correctly.

This process is only required for paper questionnaires. Online returns will automatically put people in the same order for household and individual questions.

In summary, six cases were found during the rehearsal with a suggested re-ordering for review. Scaling this up to the full census, this would suggest there could be around 700 cases for review, and around 100 cases to correct.

Though this need for name re-ordering may only affect a few cases in the overall census, it helps with the data quality for further data processing stages, resulting in less work to fix data quality issues later on. Therefore, there is a benefit to adding this processing into the overall Statistical Methods and Data Processing steps. As with other processes, further work on the sequencing with the data processing system, outputs and inputs is required to understand how the processes will run from end to end during live census.

3.5 Edit and Imputation (E&I)

Edit and Imputation fills in the blanks where individual questions have not been answered and tidies up inconsistencies between answers.

Editing is the process of locating and flagging missing, invalid and inconsistent values in the census dataset. It consists of Hard Edits and Soft Edits.

> Hard Edits**:** Rules which identify and flag impossible or highly implausible values or combinations of values. For example:
> *"A person under the age of 17 cannot drive to school or work"*
> While there may be genuine cases matching these conditions, the vast majority of observed cases will be due to error and we will want to remove these cases from the dataset. However, these edits should be documented and published for data users.
>
> Soft Edits**:** Rules which identify and flag plausible but rare observations – for example, numerical outliers such as a person under the age of 45 being retired. These are observations which we do not want to remove from the dataset, but we also do not want to multiply these values during the imputation process. These edits should be documented and published for data users.

Imputation is the process of replacing flagged missing, inconsistent and invalid responses with substituted values. There are two levels of imputation:

Item imputation: A variable-level adjustment of real households, household persons, or communal establishment persons following the edit of the dataset.

Unit imputation: The addition of a synthetic household, household person, or communal establishment person into the dataset to account for individuals and households not counted in the census.

When we mention “edit and imputation” in methodology for Scotland’s Census 2022, we are referring to item-level edit and imputation. Unit imputation is referred to as Coverage Adjustment (section 3.6).

Item level imputation may involve a limited number of deterministic changes, which determine the likely substitute value directly using other values in the dataset. However the majority of item level imputation includes an element of randomness, using a method called nearest-neighbour hot-deck imputation. This process finds similar records in the census dataset to donate values to the flagged record in order to resolve inconsistencies or replace missing or invalid values.

As part of the rehearsal, work was undertaken to see if administrative data could help with the quality assurance of the date of birth variable, which is a main census output variable. Date of birth is used to calculate age on census day, which is an important predictor variable for the edit and imputation process as well as one of the main output variables.

Two different processes on date of birth were applied as part of the rehearsal:

1. to help deal with cases where the date of birth was missing from the census
2. where there was a different date of birth on the administrative data source than the date of birth given in the census

Concerning cases where the date of birth was missing from the census questionnaire, these were linked to the administrative data source. Around two per cent of records in the rehearsal dataset had a missing date of birth.

About half of the missing date of birth records could be found in the administrative dataset for rehearsal. The administrative data date of birth could then be used to assist donor imputation in the Edit and Imputation process. This improves the suitability of the census donor record which benefits the data quality and subsequent outputs being produced by the census.

As part of the process for checking date of birth, we identify where date of birth is different between census and administrative data. For example, this can be an issue on a paper questionnaire where the information has been mis-scanned, possibly due to figures looking similar when handwritten (e.g. 6 and 8). This administrative data check will allow us to quickly identify scanned paper questionnaires that need to be manually checked.

Administrative data can also be used to help identify records where under-16s appeared to be adults with an incorrect year of birth, for example, writing the current year 2022 instead of their date of birth year. An automated process for checking these records would use administrative data along with other information from the census record such as marital status, employment and qualifications. This remains in development, so this process was not tested during this rehearsal.

Questions addressed during rehearsal	Key findings	Next steps
Is the data supplied in the correct format?	Data transformation was required to allow the data to be processed through E&I.	Further work is being undertaken to ensure updates are made and the data from live census will be in the format required for data processing.
Does the E&I process complete successfully?	Yes. The process was able to run successfully and output results.	Further development of E&I code to ensure it is ready for live census is currently underway.
Are the results of sufficient quality?	Respondent behaviour for the rehearsal is not indicative to that of a census, as the rehearsal is voluntary and the census is not. This means statistical quality is difficult to assess. However, the data was able to be processed successfully, with the data transformations mentioned above.	Work on developing the quality assurance methodologies and a quality assurance dashboard for this process is underway
Do the benefits of using administrative data outweigh costs?	Yes, it improves the placement of census donor records for missing date of birth. It helps identify mis-scanned paper records faster where the administrative date of birth is different from the census date of birth.	For improved quality assurance for census 2022, to run date of birth checking processes with administrative data linking and clerical review. As with other processes, further work on the sequencing with the data processing system, outputs and inputs is

		required to understand how the processes will run from end to end during live census.

3.6 Estimation and Adjustment (E&A)

The census aims to capture details of the whole population of Scotland. However, it is expected that during the census some people and households will be missed. In addition, some people may get counted in the wrong place or more than once. Census coverage adjustment, or Estimation and Adjustment (E&A), is finding out how many households and people we collected information from during census data collection, working out how many we think there should be, and then adding in what is missing.

The estimation process requires a clean and complete census dataset and it also needs a clean and complete set of data from a second survey – the Census Coverage Survey (CCS). The CCS is a follow up sample survey of approximately 45,000 households carried out six weeks after census day. The CCS is voluntary and contains only key social and demographic questions from the census. Having this second set of collected data is key to being able to generate population estimates.

The people and households in the census are linked to the CCS to work out:

- how many were captured in both,
- how many were just found in the census,
- how many were just found in the CCS.

Using standard statistical methodology called Dual System Estimation (DSE)[6], the results of the linking can be used to estimate how many households or people were included on neither the Census nor the CCS.

Once the final estimates have been agreed from the estimation process, the adjustment process creates new records. These records represent the people and households that were missed in the census. The final census dataset should then match, as closely as possible, the estimates we have produced during the estimation process.

There are a number of ways to place these records:

- Add people to existing households

[6] More detailed information on DSE can be found in the Estimation and Adjustment Methodology paper. https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

- Add people to existing communal establishments
- Create new households in a 'space' we already know about (for example a known address that is an occupied property from which we received no response)
- Create new households in a 'space' that we don't have an address for – but allocate them a real postcode so we know where they are

This is a complex process that requires a combination of different statistical methodologies to calculate, for each census record we received, the probability that a person or household with those characteristics would be missed from the census. This information is then used to pick a number of existing person records to use as donors, with key characteristics from these donors used to create new person records – this is the unit-level imputation process.

As there was no CCS as part of the 2019 Census Rehearsal, artificial datasets had to be created in order to test the Census-to-CCS linking process and the E&A processes.

To test the linking, CCS datasets were created from an administrative data source, taking a random sample of postcodes from the areas included in the rehearsal. This synthetic CCS was created using all the records from the latest available cut of the administrative data source (i.e. as of 30 June 2019) at these sampled postcodes. Though the timescales were not exact between the rehearsal data and this CCS data, the main purpose was to test and quality assure the linking methodology.

The main reason for doing this exercise in the rehearsal, was to check that the linking methodology works as expected. For 86 per cent of rehearsal returns that were in the synthetic CCS areas, a link to the synthetic CCS was strong enough to be accepted without clerical review. The remaining synthetic CCS records were clerically reviewed against lists of rehearsal records. When this linking process was tested on the census 2011 data, about 91 per cent of the 2011 CCS linked to the 2011 census with an automatically accepted link. The difference could be due to when the snapshot of the administrative data was taken (3.5 months ahead of rehearsal).

To test the E&A processes, another version of the CCS was created by copying the rehearsal data. The low response rate in rehearsal meant that the CCS generated from administrative data had a linkage rate too low for the E&A processes to function and therefore be tested. In order for the rehearsal data to have "Census only" records, and this CCS to have "CCS only" records, some records were removed and new records were added.

Questions addressed during rehearsal	Key findings	Next steps

Is the data supplied in the correct format?	Data transformation was required to allow the data to be processed through E&A. Some data sources - CCS, placeholder data, alternative household estimate for Household Bias Adjustment and overcount propensities - were produced synthetically as they were not available for rehearsal. To test the linking process, CCS datasets were created from an administrative data source. To test the E&A processes, the CCS was created from the 2019 rehearsal data itself. The synthetic CCS produced from either administrative data or 2011 census data did not have a high enough match rate for the E&A processes to function and therefore be tested. Identifiers for individual and household records had to be created.	Further work is being undertaken to ensure updates are made and the data from live census will be in the format required for data processing. Further work is required to establish how unique identifiers for individuals and households will be created in live census.
Does the E&A process complete successfully?	Yes. The process was able to run successfully and output results.	Further development of E&A code to ensure it is ready for live census is currently underway. Some E&A methodologies have still to be finalised prior to live

		census, e.g. household bias adjustment[7]. This work is currently underway.
Are the results of sufficient quality?	Respondent behaviour for the rehearsal is not indicative to that of a census, as the rehearsal is voluntary and the census is not. This means statistical quality is difficult to assess. Also, the use of a synthetic CCS makes quality more difficult to assess. However, the data was able to be processed successfully, with the data transformations mentioned above.	Work on developing the quality assurance methodologies and a quality assurance dashboard for this process is underway.
Have we identified problems which require a change to other statistical processes?	No major issues were identified, although E&A needs to be aware of names of flags[8] within previous processes, particularly RFP and E&I.	Further work is being undertaken to test that all the statistical methods work as expected when they are run together, as they will be in live census.
Have we identified problems which require a change to CCS collection?	The rehearsal raised the question about what the format of non-response information for CCS addresses would be.	Further work is being undertaken to ensure updates are made and the data from live CCS will be in the format required for data processing.
Did the administrative data linking methodology run successfully?	This rehearsal provided confidence in the methodology for matching the census and CCS. The linking processes was	As with other processes, further work on the sequencing with the data processing system, outputs and inputs is

[7] The probability of a person responding to the CCS should be independent of whether they responded to the Census. If this independence does not hold, it causes bias in the population estimates that are produced. Household bias adjustment is part of the Estimation process which identifies the level of dependence between the Census and the CCS and corrects the bias introduced by this dependence accordingly.

[8] As part of the flag strategy mentioned on page 13, steps further along the processing chain such as Estimation and Adjustment will need to know what earlier processes call their "flags" (variables that are used to search for records of concern) in order to address the concerns raised by them.

	manually checked (clerically reviewed) for every link (automatic and manual) in the rehearsal, to check that the methodology was working correctly. Of those links which were passed for manual review, around 1.5 per cent of these were found to be matches.	required to understand how the processes will run from end to end during live census. Refinement of clerical review process and how records are passed for automatic and manual review.

3.7 Census De-duplication

The rehearsal tested methodology that was developed to de-duplicate the census. This methodology is used to identify where a person may be recorded in more than one household in different locations. This is different to RMR, as RMR deals with duplicates within the household or within a postcode.

To do this, the census is linked to itself to try to find any instances where someone may be recorded twice. If people who are recorded on more than one census return are not identified and accounted for then this will lead to an overcount in the population.

In 2011, only links where the name and date of birth agreed exactly were considered a duplicate, and there was no checking against administrative data. Linking was only done on a sample of census records to calculate what proportion of records were overcounted. These proportions were then used in DSE to correct the population estimates accordingly.

For 2022, a method was developed to mitigate this potential overcount and was tested during the rehearsal. Once the census dataset has been linked to itself, the linked records are then linked to an administrative data source. If both of the linked census records have corresponding administrative data records at their locations then this is taken to indicate that the census records represent distinct individuals. For the remaining records, a probability score is used to calculate the chance of two census records being the same person. This information is then used in the Estimation and Adjustment process to estimate the population as accurately as possible.

The rehearsal results show that there is a benefit to using this process as part of the overall Statistical Methods and Data Processing steps. As with other processes, further work on the sequencing with the data processing system, outputs and inputs is required to understand how the processes will run from end to end during live census. There is also work to do to decide how the probability scores will be used in the Estimation and Adjustment step.

4. Conclusion

The Statistical Methods and Data Processing Rehearsal was an important step in testing and developing the statistical methodologies and data processing steps for Scotland's Census 2022.

The results showed that the processes were able to run successfully and highlighted areas which required updating or further developments to be made before the census goes live in 2022.

Work will continue through to 2022 and beyond, to ensure that we are ready for the census. This includes working through the improvements identified by the rehearsal, finalising the census statistical methodologies and how the processes will be run. We will continue to test our methodologies and the systems for running data processing, working closely with our counterparts across the UK, ONS and NISRA, to ensure our methodologies are harmonised as much as possible and learning from their experiences in 2021.