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

Methodology Enhancements to Secure High Quality Census Outputs and Population Estimates

June 2023

Foreword from the International Steering Group

The International Steering Group (ISG) was set up in May 2022. It is reporting its findings to the Registrar General (RG) at National Records of Scotland (NRS). The focus is to support the RG and her teams to construct a database that will be used to provide the outputs from the Census.

The ISG has focussed and steered work done by NRS to look at lower than expected return rates. The ISG has advised on the use of two established tools that combine with the census responses to make up a modern census. The first is the 'Census Coverage Survey'. This is a survey which helps assess how good the data collected during the census is. The second is administrative data. These are data, gathered by public bodies, that the team at NRS is allowed to access and can help to improve the quality of the census database. The latest findings of the ISG, outline how the work NRS has done is in line with other modern censuses. It also details how NRS has switched to a method called 'direct modelling'. Such an approach is used by ONS and other countries.

This is the third statement that the group have made. The full text from ISG reads below:

"As outlined in the methodology update released by NRS, modern census processes utilise the combination of a main census data collection, a census coverage survey (CCS), and administrative data to produce robust statistical estimates and outputs. Therefore, it was always planned that the coverage survey would be used to measure and adjust for census non-response. Following the close of fieldwork for both the main census and CCS, focus shifted to evaluation of the proposed methodology and administrative data options to enhance that methodology, to confirm methods to be used in the statistical processing phase.

Administrative data option evaluations indicated augmenting the CCS response data would be the best way to incorporate high quality administrative data for non-responding households into the estimation process. This would add confidence to

the production of accurate census estimates. The research demonstrated that this provided more benefit than using administrative data to augment the census response data. The decision to augment the CCS data also prevented the need for significant change to the broad population estimation approach. The switch to use of direct modelling for coverage with logistic regression rather than the planned implicit modelling approach of 2011, while a change, brings NRS in line with the 2021 approach implemented by ONS and international approaches using direct modelling.

ISG have also provided advice to NRS on several other methodological decisions outlined in the paper, which will ensure there is a solid foundation to produce the census database.

These decisions were made on a sound methodological basis, providing NRS with a strong position for the statistical processing phase."

Contents

1. Introduction 5
2. Scotland's Census 2022 5
3. Statistical Methods 6
   3.1 International Steering Group 7
4. Changes to Approach to Secure High Quality Census Outputs 8
   4.1 Coverage Estimation Planned Published Approach 8
   4.2 Enhancing Coverage Estimation 10
   4.3 Coverage Adjustment Planned Published Approach 15
   4.4 Enhancing Coverage Adjustment 15
   4.5 Quality Assurance and Validation Planned Published Approach 16
   4.6 Enhancing Quality Assurance and Validation 17

1. Introduction

This paper provides a summary of the changes that the National Records of Scotland (NRS) is making to its approach to securing high quality Census outputs and population estimates. These changes have been made to ensure that Scotland's Census 2022 delivers the benefit required by its many users. Statistical methods were already in place to produce estimates for the whole population (rather than just who completed a questionnaire). The use of some of these established methods have been revisited and updated in response to lower than planned for return rates from the collection phase.

In September 2022 NRS published an overview of the areas of research it was intending to take forward following the data collection phase of the census. This included ground-breaking work to combine information collected in the census itself, the Census Coverage Survey, and administrative data.

Our research has been guided by an International Steering Group (ISG). The ISG was established by the Registrar General to provide external advice on the quality of the census and the statistical approaches that will be used to produce high quality census outputs.

2. Scotland's Census 2022

Statistics from Scotland's Census will be the most complete source of information about the population and households. All households in Scotland were asked to complete a questionnaire during the main 2022 Census collection phase which took place between 28 February and 1 June 2022. Individuals who were not living in households (for example in a student hall of residence or care home) were also asked to complete a questionnaire.

No census achieves 100% coverage of the entire population. Since the 2001 Census, we have used internationally recognised statistical methods common across the

United Kingdom to deliver estimates which do represent 100% of Scotland's population. Statistical methods use data collected in the census in combination with the Census Coverage Survey (CCS) which was carried out immediately after the main census collection period.

The design of the statistical methods was independently assured prior to the collection phase. It was this design which defined a target of around 94% of households nationally completing a census questionnaire and a minimum of 85% completing a questionnaire across Local Authority areas.

The final household return rate from Scotland's 2022 Census was 89.2%. Only two Local Authority areas were marginally below 85% at 83.3% and 84.0% respectively, with 19 out of 32 areas achieving a return rate of more than 90%.

In September 2022 we published a summary of where we could adapt the original design through innovative use of administrative data to ensure the robustness of census statistics. This paper sets out the decisions that have been made following research carried out under the guidance of the International Steering Group (ISG).

3. Statistical Methods

A summary of the data journey for producing census statistics was published as part of the September 2022 publication. This included the capture, coding, cleaning of data and processes that ensure that the information provided by the public is complete and consistent at record level. Changes have not been required to these stages.

NRS has focussed research on how to enhance the coverage estimation/adjustment and quality assurance/validation stages of the data journey.

Changes have also not been required for the final stages of the data journey, Statistical Disclosure Control (SDC) and the publication of outputs. Innovative

methods for these later stages were already planned for Scotland's Census, building on international best practice. SDC processes will be applied to the final dataset to ensure that individuals and households cannot be identified in published statistics. Through the development of a Flexible Table Builder, we will enable users to generate the statistics they are most interested in with detailed multivariate data being available more quickly than the equivalent from the 2011 Census.

Further information about our statistical methods can be found on Scotland's census website.

3.1 International Steering Group

The International Steering Group was established in May 2022. The group's role in Scotland's Census 2022 is to:

- provide assurance that the census programme was ready to move on from the collect phase.
- provide direction and support to National Records of Scotland as we implement our statistical design after the collection phase
- propose amendments or changes for us to consider, including accessing additional administrative data.

ISG is chaired by Professor James Brown, ABS Professor of Official Statistics at University of Technology Sydney and is made up of pre-eminent international authorities in census coverage and use of administrative data.

In June 2022 the ISG stated that Scotland's Census 2022 has a 'solid foundation' on which to build as we prepared to move forward to the Census Coverage Survey. ISG also stated that 'the coverage survey, combined with innovative use of administrative data, will allow NRS to estimate the size, shape and characteristics of the population as planned'.

4. Changes to Approach to Secure High Quality Census Outputs

4.1 Coverage Estimation Planned Published Approach

As stated in section 2, counts of who responded to the census are used in combination with the responses to the Census Coverage Survey (CCS) to produce estimates covering the whole population. The approach relies on very careful matching between those who responded to the census and those who responded to the CCS. These links are used to determine who responded to both the census and CCS as well as who responded to one but not the other.

Dual System Estimation (so called because we use two sources in the method) is then used to provide an estimate of how many households or people the census missed. For example, if 90% of those responding to the CCS also responded to the census, then we would estimate that 10% of the population did not respond to the census.

The CCS is the second largest social research exercise in Scotland after the census itself. Instead of covering all areas, the CCS samples all households in selected postcodes. Around 1.5% of all households are included in the sample. Dual System Estimates are calculated across sampled postcodes with a process of Ratio Estimation used to determine population and households estimates for each Local Authority area.

As with any estimate there is a degree of uncertainty with the statistics produced. We still plan to measure this uncertainty through producing and publishing confidence intervals. There is also a risk that final estimates could contain an amount of bias if the assumptions that underpin the statistical methods do not hold. A key assumption is that the census and CCS are independent, that the probability of responding to one is independent from responding to the other. Heterogenous response rates (where there is wide variation across the country for example) can also cause bias.

A comprehensive quality and validation exercise was a fundamental part of our planned design. Similar exercises were carried out for the census in both 2001 and 2011. By analysing the coherence between census estimates and other available data we will consider whether there is evidence that our estimates need to be revisited or whether a lack of coherence reflects uncertainty with an alternative source. One of the most important reasons for conducting a census for example is to re-base the Mid-Year Population Estimates (estimates of the population produced in between census years) which will contain inaccuracy from how population change is measured annually, mainly due to the challenge of measuring migration.

The planned coverage estimation approach for Communal Establishments (CEs) differed depending on the size of the establishment, following a similar approach used for the 2011 Census. CEs are residential accommodations that are managed, for example hospitals or hotels. Small CEs (with fewer than 100 bedspaces) were included in the CCS so we planned to use a similar Dual System Estimation approach as with households. We did though intend to produce estimates across the whole of Scotland rather than separately for individual Local Authority areas. In addition, we boosted the CCS sample in areas with small CEs.

Large CEs were not included in the CCS due to the time and resource required for successful enumeration. Instead, we planned to use additional information that we collected from the Census CE Managers' Questionnaire which asked about the number of male and female usual residents each establishment had, broken down into different age groups. This additional information is similar to an approach used by the Northern Ireland Statistics and Research Agency (NISRA) for the 2021 Census in Northern Ireland. Coverage estimation would then be a manual process using this information.

Further information on the original coverage estimation design for the household population is available on our website along with further information on CEs. We have also published details of the planned approach for the validation of population estimates.

4.2 Enhancing Coverage Estimation

Research has been guided by the International Steering Group. We publish notes of the regular meetings on our website.

Decision 1 – Use of a National Logistic Regression Model

We published details of our intention to change the design of our coverage estimation approach in our September 2022 update. Rather than using three stages (dual-system estimation, ratio estimation and synthetic Local Authority estimation), we are using an approach based on a single stage logistic regression. This is the same approach used by the Office for National Statistics (ONS) for its 2021 Census. As with the previously planned approach, census and CCS records are still the basis of a Dual-System Estimation approach. The logistic regression approach uses a single national model to calculate record level non-response weights which are used to estimate council population and households. The model in effect predicts the likelihood of a person responding to the census accounting for age, sex, ethnicity, and other potential variables.

Decision 2 – Supplementing Coverage Estimation Calculations with Administrative Data

We also stated in our September 2022 update that we were evaluating whether to include administrative records in our coverage estimation calculations. We have assessed whether to use such records as part of either the census or CCS in place of people who did not respond. Our research has concluded that supplementing the CCS provides the greatest benefit in the overall quality of the statistics. Administrative records are used to enhance coverage calculations then removed from the CCS in-line with commitments made with data suppliers.

Our assessment was based on two considerations:

a. Is there evidence of extremely low localised response rates in the census?

b. Which approach is likely to lead to the greatest decrease in the level of uncertainty in the final estimates?

Consistency (or homogeneity) in return rates is an assumption underpinning our coverage methods. We set targets both for the total return rate and for consistently high return rates across the country. If there was evidence of particularly low return rates at neighbourhood level it would be preferable to supplement the census with administrative records.

NRS has re-evaluated the level of homogeneity/consistency and did not find evidence of low or inconsistent neighbourhood return rates which justified supplementing census data directly.

An evaluation of the decrease in the level of uncertainty, by supplementing either the census or the CCS, was based on a series of simulation studies. This involved modelling with representative synthetic data and measuring uncertainty through confidence intervals around the resulting estimates. We concluded that supplementing the CCS provided a greater reduction in uncertainty than supplementing the census itself.

Decision 3 – Selecting Administrative Records to Supplement the CCS to Enhance Coverage Estimation Calculations

While supplementing the CCS with administrative data in coverage estimation calculations can decrease the level of uncertainty in the estimates, we recognise it can also lead to error if an administrative record is out of date. An administrative record may be out of date, if for example, an individual has moved house recently but not yet updated their health provider. This would tend to result in an overcount if an individual on an administrative source was no longer resident at a CCS address at the time of the 2022 Census. We therefore took a conservative approach to which records to use to limit any overcount.

We have been working closely with administrative data owners over recent months to ensure we have access to the sources identified by the International Steering Group as 'essential'. Access has been secured for all sources regarded as essential for use in the 2022 Census.

Decisions about which administrative records to supplement the CCS with were based on a careful consideration of the 'strength of evidence' from administrative records. There will be greater confidence in the accuracy of a record if the same address information is confirmed by multiple sources.

We took a data driven approach to determine the optimal set of records to use. Beginning with an initial administrative data set, we evaluated the extent to which applying restrictive rules to the administrative records (for example a rule that a person needed to appear on multiple datasets or have recent interactions with the data provider) removed records. We then considered the extent to which there was evidence that the additional rule was removing the records we were interested in (hadn't responded to the CCS or census) and hadn't removed records that we were confident in (because they had responded to the CCS or census).

As expected, we found that appearance on multiple administrative datasets was stronger predictor of someone being in the Scottish population. This was stronger than an indication from a single source such as recent health activity. ISG recommended that we only supplement the CCS with administrative records we are very confident in as being part of the Scottish population.

The administrative sources to be used vary depending on the age groups cover but include:

- Birth registrations
- National Health Service Central Register (NHSCR)
- Health Activity (individuals who have interacted with selected NHS services within the last 3 years)

- School pupil census
- Electoral register
- Higher Educational Statistics Agency (HESA) data

Decision 4 – Use of Administrative Data in Coverage Estimation Calculation for Communal Establishments

In re-evaluating the role of administrative data in coverage estimation calculations for households, we have also re-evaluated how such data can be used for coverage estimation calculations for people in CEs. We considered approaches for using these data alongside the additional information collected through the CE manager form (listing residents age/sex characteristics) and information collected in the CCS. Administrative data are only used to inform the age/sex distribution and total number of residents in CEs.

The source used for coverage estimation will be dependent on CE type and quality of administrative data available. CEs contain a range of establishment types so for each we have prioritised the following:

1) Trusted Administrative Source. Where we have confidence that records are kept up to date with who is resident. This includes use for both prisons and military bases where we were unable to collect an age/sex listing.
2) CE Manager Form Age/Sex Breakdown. Where administrative records are not available or are not updated frequently and where we are likely to have stronger evidence from the CE Manager Form. This includes use for student halls of residence.
3) Linked data spine. Where nether a CE Manager Form or a single administrative source is available or where they look inconsistent with other evidence for a particular CE, we will use our 'linked data spine' which is based on a range of sources including each of those listed for the previous decision on supplementing the CCS.

Within institution types we will make an assessment on a case-by-case basis. Triangulation will be carried out using the sources listed above as well as the count of usual residents on the CE manager form and the information on the number of bedspaces on our pre-census collection list of CE addresses.

Decision 5 – Use of an Alternative Household Estimate for Bias Adjustment and QA

As described in September 2022 we intended to develop an Alternative Household Estimate (AHE) as an additional source to identify and potentially correct for bias in our household coverage estimation process. Our work draws heavily on the methodology developed by the ONS, estimates from which proved a valuable additional quality assurance tool for them.

The AHE is a distinct and separate method for determining the number of occupied households in Scotland to the coverage estimation approach described in previous sections of this paper. It brings together responses to the census, the original address frame, administrative data, and information collected during the census field operation.

All households which returned a census form and stated that the address was occupied by at least one usual resident are included in the AHE. These responses make up the vast majority of the estimate.
We marked non-responding addresses as vacant where if an address had been visited by our field operation at least four times and was identified as 'probably unoccupied'. Non-responding addresses were also marked as vacant if they had been visited at least three times in the field, had been identified as 'probably unoccupied' by a field officer and had no sign of activity on administrative data.

For the remaining non-responding addresses which did not match any of the criteria above we used a combination of information including field intelligence and administrative data. We were able to use an indicator of unoccupied addresses based on Council Tax information collected from all Local Authorities as a

particularly important source with discount and exemption information used as a determinant of occupancy.

4.3 Coverage Adjustment Planned Published Approach

At the end of the coverage estimation phase, we will have estimated the total population and households at Local Authority level by some key characteristics such as age, sex and ethnicity. The adjustment phase then allocates people and households to record level so that the census database from which all statistics are produced has complete coverage.

Adjustment adds creates new records by either:

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

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

4.4 Enhancing Coverage Adjustment

Decision 6 – Use of Administrative Data in the Placement of Non-Responding Persons

Our research has concluded that we can improve our initially planned approach by using administrative data to improve the accuracy of donor records on which to adjust the database for under coverage. Where we have a person in the

administrative data that did not respond to census, we can use the information about their characteristics (age and sex) to better select a similar donor in the census dataset. We can then use administrative records as markers for where these new records should be placed, taking into account the geographical location, age and sex of these records to place a similar record in the correct location. Using the administrative data in this way will allow us to place the right records into the correct addresses and reduce the chance of us placing records into vacant households rather than non-responding households.

We will use this admin data in this way to adjust for a portion of non-response. Once this has been completed the rest of the deficit between our census population estimates and census returns will be made up through an adjustment algorithm similar to what was previously planned. The algorithm will select responding households from census that are similar to those who are still missing. Key characteristics of these households, and the people in them, are used to create new records. These records will be placed at an address which did not respond and a combination of our fieldwork information and administrative data suggests is occupied.

4.5 Quality Assurance and Validation Planned Published Approach

The changes to the methods for producing census estimates do not fundamentally change the way in which we will quality assure and validate the data before publication. Our published Statistical Quality Assurance Strategy made a distinction between quality assurance at every step of the census data processing journey and validation of final estimates.

By assessing quality at each step, we can confident that we are not introducing errors while we process the data. As described above, data validation is an assessment of the coherence between the final estimates and other available data. We will consider whether there is evidence that our estimates need to be revisited or whether a lack of coherence reflects uncertainty with an alternative source.

4.6 Enhancing Quality Assurance and Validation

Decision 7 – Enhanced Review of Links Made Between Census, CCS and Administrative Data

We have extended the 'Assurance of Processes' to cover the changes we've been making to coverage estimation. Additional checks have been carried out on all links between administrative data and census and CCS records. This involved manually reviewing samples of automated links to have confidence in the outcomes. We also analysed links at aggregate level to check for any unexpected patterns. These checks allow us to be confident that we are only adding administrative records to the CCS for individuals not already in the CCS and be confident about whether each administrative record responded to Census or not.

Decision 8 – Collaboration with Local Authorities in Validating Census Estimates

We have also enhanced our approach to validation of census statistics. All Local Authorities were asked to provide us with Council Tax data as an additional source for QA as well as for the Alternative Household Estimate. Local Authorities were asked to use exemption and discount code information to be able to identify properties which were likely to be occupied and vacant. This data will be used to supplement the other sources we will use for validation as set out in our published plans. This includes Mid-Year Population Estimates, estimates from the 2011 Census, Scottish School Pupil Census, National Health Service Central Register, and data from the Higher Education Statistics Agency. More information on our Validation of Population Estimates methodology is available on the Scotland's Census website.

As in 2011 where we included Local Authorities in quality assuring census population estimates we continue to recognise the value that Local Authorities play in validating the estimates. With this in mind, we will send participating Local Authorities early census estimates for their area, which may yet change before publication, for their area to compare to their own locally held data. Any

inconsistencies identified will be evaluated as part of our validation process to determine whether there is a need to revisit our methodology. More information on our plans for quality assurance panels is available on the Scotland's Census website.

Sharing estimates before publication is permissible under the Code of Practice for Official Statistics but for the purposes of quality assurance only.