Census 2021 data processing overview

Data processing for Census 2021 is based on methods successfully used for Census 2011. Census Office have worked closely with other Census offices in the UK, Ireland and internationally to learn from best practice and ensure that data processing methods are well-designed, reliable and statistically robust.

The Census estimates will be adjusted to account for estimated non-response using information from the follow-up census coverage survey. Census 2021 outputs will therefore represent our best estimate of the whole population of Northern Ireland (NI), not just those who responded.

The following are the main stages of data processing for Census 2021:

- Data consolidation
- Data transformation
- Coverage
- Preparation for Outputs

Quality Assurance

Data Consolidation

As most of the response data for Census 2021 will be collected online, the expectation is that data quality will be higher than in previous censuses. The broad aim of the data consolidation process is the removal of invalid records and responses, including duplicates. This involves two main stages, namely remove false persons (RFP) and resolve multiple responses (RMR). Census records that do not include enough information to be considered acceptable responses are identified and removed via RFP. The RMR process consolidate duplicate responses at each address. However, for 2021, the RMR process has been extended to search and resolve duplicated persons/households across the entire population.

Another innovative 2021 step, is the management of returns on students. For 2021 (as in 2011), students were required to be enumerated at their term-time address to satisfy the usual resident population base. However, due to the known issues associated with maximising the student enumeration, and the prevalence of COVID restrictions, it was decided to allow students to also be enumerated at their home address and ask what their term-time address was on the questionnaire. This information will be used within the RMR process to move students to their term-time address in the database.

Data Transformation

Data Transformation is the process that makes edits to the Census responses due to missing, invalid or inconsistent data. As was done in 2011, rules will be applied to census responses to ensure that questionnaire routing has been followed, and to remove implausible and impossible variable relationships from the census data. For example, to correct issues such as labour market questions being answered for children aged under 16 years.

The edit and imputation process identifies erroneous data, and where necessary it imputes a value based on the likely response. This will ensure a complete and consistent dataset for subsequent processing. As for Census 2011, the Canadian Census Edit and Imputation System (CANCEIS) will be used. This specialist program is well-established and has been successfully used for the edit and imputation of censuses in many countries; it employs a donor-based, minimum change imputation approach.

Coverage

The Coverage process in 2021 will again repeat the process used in 2011. As with the other UK Census offices, the Census Coverage Survey (CCS) will be used to estimate the level of undercoverage in Census 2021. The CCS gathers data from approximately 15,000 households in a specific sample of postcodes across NI. CCS data will be linked to the corresponding Census data where available to determine the extent of undercoverage; this information will be used in a dual system estimation method (commonly known as capture and recapture) to generate population estimates for both sampled and non-sampled areas.

Coverage adjustment will amend the Census dataset so it reflects the population estimates derived from the coverage estimation process; it will therefore account for people and households that were missed in the Census.

Preparation for Outputs

The steps involved to prepare the data for outputs are broadly the same those adopted in 2011. They are:

- Add geographies
- Statistical Disclosure Control
- Apply Derived Variables for Outputs (DVOs)

The step to add geographies takes every response and attaches an output geography to it to allow for aggregate outputs at local geography and statistical geography levels. Once complete the dataset is passed through the Statistical Disclosure Control process where statistical 'noise' is introduced to protect against any individual or household being identified. Finally, a number of DVOs are calculated from the response variables which combine response information to create variables that are required for output tables. For example, Employment Status, National Statistics Socio-economic Classification, Living Arrangements, Household Composition etc.

Quality assurance

Each stage of data processing will have bespoke quality assurance checks. Examples are likely to include:

- Clerical checking of records flagged as multiple responses during data cleaning
- Checks during edit and imputation to ensure that particular donor records have not been overused
- Socio-demographic analysis of the data after coverage estimation and adjustment to identify any anomalies against broad expectations

- Spatial and time series analysis of the data to ensure trends look plausible and where they are not, an explanation can be given (for example, effects of COVID pandemic on workplace location)

The main aim of the quality assurance checks is to ensure that the Census estimates are plausible at the NI level and also at lower levels of geography such as Local Government District and Small Area. Further details on the quality assurance for Census 2021 are available in the published 2021 Census Quality Assurance Strategy.