Scotland's Census 2022

External Methodology Assurance Panels

Summary Note PSR005: Panel 5

Wednesday 23 September 2020

Contents

PSR005: Summary Report of the findings of EMAP Session 5 – Wednesday 23 September 2020

1. This paper summarises the main points of discussion during the external methodology assurance panel, including overall conclusion and advisory recommendations.

2. Where appropriate, the panel's reasons for any advice that proposed methodology is not fit for purpose will be stated.

3. This paper will be published on the Scotland's Census website, following approval by the panel.

4. The methodology papers reviewed by this panel were: -

PMP012: Overview of Edit and Imputation for 2022

PMP013: Item level Checking using Administrative data – Date of birth methodology

PMP014: Resolving Multiple Returns methodology

Head of Statistical Quality Assurance team
Scotland's Census 2022
National Records of Scotland

Email: scotlandscensus@nrscotland.gov.uk

1. PMP012: Overview of Edit and Imputation for 2022

Main points of discussion:

The purpose of this paper is to give an overview of the Edit and Imputation methodology. The Edit and Imputation process focuses on detection and correction of missing and inconsistent responses.

Donor imputation is the main method used in Edit and Imputation. This involves finding similar records in the census database for each record that needs to be fixed, and using the information from the donor record to input blank records or correct inconsistent responses. The Canadian Census Edit and Imputation System (CANCEIS) is the software used for this process.

The paper outlines the changes and improvements in the methodology from the 2011 Census, including modularisation, using processing units, imputation of partial codes, and improvements to the imputation of relationships which were previously imputed using a deterministic algorithm.

The paper provides details on further changes and improvements to the methodology including changes to the edit rules, the use of admin data for imputing age and student status.

1.1 The panel recognised the quality of the paper. They thought it was well written, the key areas clearly defined and the information on improvements from 2011 easy to follow. The panel praised the quality of the methodology, and pointed out that the paper sometimes understates the huge amount of work and expertise in NRS to produce this work.

1.2 The panel was content with the approach and agreed that the methodology was appropriate and very robust. Most of the comments and discussion focused on adding some additional detail to the descriptions of the methods.

1.3 The panel praised how the previous methodology had been developed to exploit the advances in computing power. It was agreed that CANCEIS was an appropriate and reliable tool, and the current methodology has pushed its functionality even further. The panel was impressed with the methodology of the Edit and Imputation process of using record level donor imputation values. The methodology was also used in 2011 and can help correct bias by looking at similar individuals in the area. It was thought that the paper would benefit form a further discussion of alternative tools and why CANCEIS was chosen specifically.

1.4 The panel questioned the potential frequency of invalid and inconsistent values in the context of the data collected mainly through the online questionnaire. NRS explained that the online questionnaire's built-in functionality is expected to minimise missing and inconsistent records through automatic prompts and error messages.

However, the approach is not intended to increase respondent burden by overwhelming the respondents with error messages. The online system includes soft validations and hard validations. The hard validations are applied to key variables only. Some of the validations are based on evaluating the results from the 2011 Census and the 2019 Rehearsal. NRS will add more information to the paper about how the online questionnaire functionality will reduce missing and inconsistent values.

1.5 There was a general discussion about the administrative data sources. The panel suggested to include examples illustrating the use of administrative data for matching and imputing for census data. Panel recognised that the methodology of using the administrative data is extremely sound. More details are required in the paper on how the differences between the census data and administrative data will be dealt with, and how the bias between data linkage and imputation is assessed.

1.6 In the discussion on modularisation the panel agreed that the grouping of modules is logical and sound, and further questioned whether variables can overlap two groups. The panel suggested that it would be useful to include more detail on how exactly these groupings work, and how the ordering relies on each other. NRS explained that the methodology and the software used does not allow for the variables to be imputed in two modules as this will create inconsistencies. However, further consideration will be given to using these variables in different modules as predictors. NRS confirmed that detailed modularisation methodology is a part of the next piece of work. NRS will add more detail about the ordering in the next methodology paper on modularisation.

1.7 There was a general discussion of the section of the paper on applying edit and imputation methodology to the voluntary questions. The panel highlighted that the question is intended to capture the data on people of minority, who by definition might chose not to respond to the question. NRS responded that in the case of voluntary questions specifically there is an important difference between outputting a count versus an estimate. The count represents how many people answered with a specific response option, whereas the estimate offers data on groups of people that add up to the whole population. NRS explained that in the case of invalid or inconsistent responses, some invalid responses are treated in the same way as missing because it is not a valid response. For example, in a long-term conditions question a write-in response that is not a long-term condition. Inconsistent responses, on the other hand, occur when there is a logical conflict between the responses to questions. NRS confirmed that there is not a lot of potential for inconsistences for the voluntary questions, and that other data will not be used to correct these inconsistent responses. There are quality assurance processes that deal with invalid responses (for example, two ticks when one is required) at the coding stage of the processing. Similarly, a voluntary question on religion was not imputed in 2011.

1.8 The panel asked whether the methodology was tested by artificially creating missing data, imputing it, and testing whether the imputed data matches with the artificially removed information. NRS confirmed that this method is already being used for testing the process methods, and a description will be added to the paper for

clarification. Similarly, further information will be included about methods of assessing the underlying trends of the missing data.

1.9 The panel suggested that research data users might prefer to work with the data with missing values, rather than imputed data. This issue is highlighted when other data sources are used, and it is difficult to identify if the data does not match due to the quality of the source, or imputing. Flagging if response is imputed would be very useful. NRS pointed out that imputation flags will not be provided for the aggregate output tables, but imputation flags will be available for record-level dataset extracts, if statistical disclosure control allows. The paper will be updated to include this clarification.

1.10 In the discussion about the processing units (PU) panel agreed with the approach of using the whole of Scotland as a dataset rather than partitioning the dataset into 10 geographically based processing units as in 2011. This approach will provide an optimal number of donor records.

1.11 The panel asked for more details on the motivation for the nearest neighbour technique to understand how it works and what trade-off are made in the data. The methodology is partly based on the data being sorted geographically. NRS briefly explained the challenge of presenting a two-dimensional map in a one-dimensional list of data. The geographical areas are numbered and ordered in a way that every record is near other records from the same geographic area. The software used in the process searches for donors by moving outwards from the failed record, so that the nearby records have the highest chance of being considered as donors.

1.12 There was a general discussion on edit rules set out for the imputation. Specifically, the relationship rules referring to half-siblings to reflect the modern demographic structure, and the age rules that might differ from the pre-defined life stages (for example, unusual cases such as person widowed at 20 years old). NRS uses soft edits that are created to involve the age variable. These are based on the population percentage from 2011 census data. This part of the methodology is a work in process, and NRS will be carrying out further research on this.

1.13 The flowchart to show where the Edit and Imputation fits in the census data processing was praised for the clarity and usefulness of the presentation of the process. It is easy to understand the order and flow. There were suggestions of including similar flowcharts in future papers where appropriate.

Conclusion: The panel were impressed with the quality of the work and paper, and were content to recommend the methodology.

Panel Advice

	Tick('✓')where appropriate
The Panel's advice is that the proposed methodology is fit for purpose.	✓
The Panel's advice is that the proposed methodology is not fit for purpose (reasons must be stated below).	
Reasons for advice (if to not proceed with proposed methodology):	

Chair: Alan Marshall

Date: 14th October 2020

2. PMP013: Item Level Checking Using Administrative Data – Date of birth methodology

Main points of discussion:

The paper considers how NRS is proposing to manage the missing date of birth (DoB) information for Scotland's Census 2022. The DoB is a key variable that is used to calculate a person's age. This information is essential for estimating the age profile of the population. The paper describes how the administrative data sources will be used to quality assure the data where the date of birth is missing or incorrect.

The method used for linking the data is the same method as that presented in the EMAP paper on linking the census data and the Census Coverage Survey (CCS) data (PMP010: Census to Census Coverage Survey (CCS) linking). The method relies mainly on postcode linking, and other variables based on a scoring system.

The method provides an extra layer of quality assurance and is used for better imputation of age in the Edit and Imputation process.

1.14 The panel agreed that the methodology is sound and a good improvement on previous methods. The paper will benefit from restructuring to assist an easy flow. In addition, the panel would like to see more detailed information on a few points in the methodology including descriptions of the variables used for matching, the scoring of these variables, choice of administrative data sources, and how small differences between the datasets will be dealt with. These details will improve readability and will assure the readers of this robust piece of methodology.

1.15 The use of administrative data sources to quality check and impute the date of birth is very sound. The panel asked why these datasets are not used directly for date of birth (either in missing, or missing and inconsistent cases). NRS replied that the administrative data records are compared with the dates of birth of all the census records, which allows the inconsistent cases to be identified. NRS explained that additional legal, ethical and privacy issues would need to be resolved to be able to directly use data from administrative sources in census outputs.

1.16 The panel raised some general queries regarding the online questionnaire and how it is intended to be used and how unsubmitted returns are used. NRS gave a quick overview of the procedures that are created for the security around the census questionnaires, and of the issue of unsubmitted returns. More detailed explanation will be added to the paper.

1.17 The panel thought that the use of scaled up examples of the high level percentage results from the Rehearsal 2019 was very useful to understand the impact of non-response for the question on date of birth. They also thought if possible that it

would be useful to see the results of the frequency of the imputed age from the administrative data sources specifically.

Conclusion: The panel were content that the methodology was sound and to recommend it for use.

Panel Advice

	Tick('✓')where appropriate
The Panel's advice is that the proposed methodology is fit for purpose.	✓
The Panel's advice is that the proposed methodology is not fit for purpose (reasons must be stated below).	
Reasons for advice (if to not proceed with proposed methodology):	

Chair: Alan Marshall

Date: 14th October 2020

3. PMP014: Resolving Multiple Returns methodology

Main points of discussion:

The paper considers the duplication of both person and household records occurring when more than one census response is submitted. The methodological process is called Resolve Multiple Responses (RMR). The duplicates are identified within the dataset and are resolved or merged into one record using a linking methodology described in the paper. The results of this record linking are then quality assured using administrative data. This paper in particular focuses on the use of linking methodology to identify the duplicates. The methodology on resolving the duplicates will be included in a further paper.

1.18 The panel agreed that there are good developments of the previous methodology and the paper describes a clear, well justified and robust methodology. Use of visual representation of the process is very useful. The panel encouraged the use of flow charts to match the narrative.

1.19 The panel questioned the rate of multiple returns in the context of the online system. NRS explained that there is a possibility of respondents completing multiple returns either by mistake (for example, a respondent completed an online as well as paper questionnaire), or if a respondent is unable to complete the questionnaire already started because they forgot their online questionnaire password. The subject of passwords is considered very carefully because of the security and protection of personal data.

1.20 The panel asked for more detailed explanation in the paper of why the overcount is considered more problematic than the undercount in the context of the census processing. If overcount might have a greater impact on matching, the paper needs to explain why. NRS explained that the RMR methodology is applied to avoid the overcount. The process improves the data quality and resolves the issue of partially completed returns. The undercount is preferred because the methodologies in place are able to deal with the undercount much more efficiently in a separate process to RMR.

1.21 Comments were raised around the groups for each category used in automatic linking resolution. More information is needed on how the key criteria for automation will be set up. NRS will add more detailed explanations to how unsubmitted returns and partially completed returns are dealt with. And what happens to the data after the linking.

1.22 The panel thought that the section on strengths and limitations was great and that the section on data linking is very well explained.

1.23 The panel thought that the description of the clerical review was very good, and that this approach shows the robustness of the methodology.

1.24 Following the review of the paper, it was suggested that the title of the paper should include 'Part 1', as further methodology of the process is planned to be published later.

1.25 The panel asked for more detail on the use of postcodes in the matching process. NRS explained that postcodes are used rather than household addresses because a household address may be different but the postcode will be the same. There is an option for respondents to type in their address in the online questionnaire, as well as include their correct address on the paper questionnaire. The addresses for census returns will be linked by the postcode, even if the full postcode is not provided by the respondents. The postcodes are used to match the multiple records to the correct geographical area. Another methodology focuses on the identification of, for example, two records of the same person in different parts of the country (for example, children with parents who live apart were included at both addresses as usual residents). The resolution of these cases will be applied at the later stages of the process.

1.26 The panel would like to see more on detailing how the methodology will be implemented in the end-to-end process. Especially, the process of resolving non-standard cases. It might be useful to see more detail about issues identified in 2019 Rehearsal in the paper, and how confident NRS is in reconciling the issue.

Conclusion: The panel were content that the method was sound and to recommend its use.

Panel Advice

	Tick('✓')where appropriate
The Panel's advice is that the proposed methodology is fit for purpose.	✓
The Panel's advice is that the proposed methodology is not fit for purpose (reasons must be stated below).	
Reasons for advice (if to not proceed with proposed methodology):	

Chair: Alan Marshall

Date: 14th October 2020