Scotland's Census 2022

External Methodology Assurance Panels

Summary Note PSR004: Panel 4

Wednesday 26 August 2020

Contents

PSR004: Summary Report of the findings of EMAP Session 4 – Wednesday 26 August 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 the 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: -

PMP010: Administrative Data - Census to Census Coverage Survey Linking methodology

PMP011: Statistical Data Processing - Removing False People Methodology

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

Email: scotlandscensus@nrscotland.gov.uk

1. PMP010: Administrative Data - Census to Census Coverage Survey Linking methodology

NRS introduced the paper and explained that it contains details of the proposed process for linking Census Coverage Survey (CCS) records to Census records, primarily focusing on person linking and subsequent dual-system estimation (DSE) used to estimate the total population. The CCS only covers around 70,000 records so the errors which can be introduced in this methodological step can have a relatively large impact on the accuracy of the resulting population estimates. To make the exercise more efficient the proposal is to compare records within blocks[1].

1.1 Main points of discussion:

1. The oral presentation and explanation of the methodology clarified some hard-to-follow elements of the paper. It was thought that the paper would benefit from redrafting to make it easier to understand the complex concepts and proposed methodology.
2. It was thought that the paper contributes to knowledge in a significant way, offering a new and innovative approach to this Census issue that could probably be flagged up more prominently. It was noted that it was commendable to develop SAS coding for these processes, replacing commercial software and making for a bespoke package. It was thought that the abstract should flag up the innovative approaches and methodologically valuable contribution to the field.
3. **Scoring and categorisation.** It was thought that the process could be explained more fully in the paper, with an earlier statement on how those are being used to arrive at strength score etc., perhaps by bringing explanations forward from Annex. Consider including a range for the accumulated scores to clarify how those scores go forward to an overall (or total) score and categorisation. Provide an explanation of where the 22 categories came from, including background, justification and discussion of options that led to this number of categories. Similarly for collapsed scores that culminated in strength scores of 0-9, and for the decision on 20 records being a critical point in the listing of possible matches during clerical review. Scoring comes across as being somewhat arbitrary – can those scores and cut-offs be justified and explained to assist the reader in making sense of the scoring processes of 'for' and 'against', and of the overall strategy being adopted.

   NRS explained that scoring is built up through an iterative process informed from using Census 2011 data. Strength scores were developed in order to count the number of records at particular strengths. All the 'for' and 'against' scores are retained to order the census records that will be added to the list of potential links. Integer strength scores allows for comparison of links, but this can be further developed in the paper to clarify. The use of '20' was on

[1] Blocks and blocking is explained in greater length in the paper.

pragmatic grounds in that this was the number of records that can typically be comfortably seen on one screen.

4. It was thought that the methodology was really good and an improvement on 2011. It was noted that the **clerical review** process was not discussed in the paper and it was wondered if it would also benefit from further consideration if not already reviewed. It was noted that it was good to address accuracy through clerical review and there were no concerns over the methodology for false-positives; but false negatives present a risk to the process. It was suggested to consider getting a second reviewer to find a link, particularly where a good link was not being picked up under position 1 or 2 (Table 4). A lot of emphasis is placed on time efficiency in the paper, where there may be scope to put more emphasis on ultimate accuracy of process. Costs of developing the code in the first place might be reported alongside the predicted hours for clerical review to provide a more realistic cost.
   It was noted that 10 reviewers were chosen for clerical reviewing, but given likely variation across reviewers there may be an argument for fewer reviewers over a longer period to improve accuracy and efficiency of process.

   NRS explained that the main concern is to get accurate results but to also get processing times down so that publication of first results is not delayed.

5. **Technical language** needs to be reviewed or explained to avoid misunderstandings. For example use of 'bias' needs careful attention, a technical point on statistical bias as opposed to differences between estimates. Similarly on use of 'minimise' where you cannot truly minimise the error or time spent but could 'reduce' [Page 16 para 3 is an example that should be reviewed in light of this statistical argument.] Again later in the paper, please check on use of 'bias' and 'relative bias'; consider using over/under estimate instead of bias.

   NRS agreed to review and noted that the term 'bias' is part of the statistics regulation QA framework that we need to explicitly tie in with. The term bias is also explicitly used in one of the Census KPIs.

6. The panel expressed a preference for **academic references** in a technical paper; replacing non-scientific sources or sources that are not 'peer-reviewed'. There is plenty published literature on the points being made that could be inserted.
   A specific example concerned description of Damerau-Levenshtein's 'edit distance' – if this is an adapted version of edit distances, describe the modification and adaptation rather than attempting to present the whole theory.
   On a similar point, greater reference could be provided to comparable work undertaken in ONS and NISRA to strengthen arguments for the adopted approach within NRS.

7. Worked examples can be useful in clarifying processes, so consider including something that can support the reader. Use of percentages would also help clarify processes in tables (Section 6) and help the reader make sense of whether this is a big number or not.

1.2 Conclusion:

The panel were content that the method was sound and fit for purpose and made suggestions to improve the readability of the paper.

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: 16th September 2020

2. PMP011: Statistical Data Processing - Removing False People Methodology

NRS introduced this paper giving general background on data cleansing required to eliminate 'false' persons so they are removed before any further processing takes place. The online questionnaire should reduce the number of occurrences because there will be fewer returns being scanned, but this will still be an issue to address for Scotland's Census 2022.

The '2 of 6' rule (where two out of six key questions are answered which suggests that this is a return for a real person) is being modified for future Remove False Persons (RFP) methodology, to be '2 of 7' key questions as a maximum comparison. For some categories it will automatically reduce to fewer comparisons (of 5 or 6) where routing decisions are made in the online questionnaire.

Use of Administrative Data is a new strand to try to validate persons who might otherwise be mistakenly removed as 'false' persons. A second new strand concerned automatic and clerical review of person names, to identify and account for 'false' names and facilitate further matching processes.

1.3 Main points of discussion:

1. The panel valued the oral presentation and found the paper clearly written and readable. A few minor typos to be addressed, issues noted on presentation of Tables, and perhaps a bit of sign posting would enhance the paper. Strengths and Limitations were very helpful.
2. The panel queried whether there a flowchart to show how the processes fit together?

   NRS explained that high level flowcharts are available and published on Page 5 of the overview document. More detailed charts can be provided if necessary.
3. No issues with the **methodology were raised**. It was noted that the innovations around Admin data were good but only offered marginal benefits. It was suggested the scale of RFP could be highlighted earlier in the paper.
4. The panel highlighted scoring issues as raised in PMP010: Administrative Data - Census to Census Coverage Survey Linking methodology paper, where **Scoring and categorisation** process needed to be explained more fully. A lot of overlap between the papers on methodology and communication of intended strategy, so comments in 1.1 (3) apply here. It was suggested that fuller details could be within the paper rather than leaving to annex.
5. The panel sought more information on **which administrative data** are being used, and how the accuracy of those data is reported.

   NRS explained that a primary data source will be NHS Central Register (NHSCR), but final confirmation of associated Data Protection Impact Assessment (DPIA) is needed before formally including data sources in this paper. Quality Assurance of Administrative Data (QAAD) sources are available

separately, reporting on quality of coverage, accuracy and relevance for the project.

6. **Presentation of Tables** needs to be reviewed for clarity of message and to restrict reporting to essential elements only. The panel recommended avoidance of variable names and to make judicious use of imagery in tables.
7. Where **privacy** is likely to be a concern for respondents, the panel queried whether partial responses on date of birth (DOB) could be used.

   NRS explained that if year and month are available then that could be utilised, but that concealed DOB would be treated as missing. NRS must be careful not to include weak records that would adversely impact on DSE processes.
8. **Blocking strategy** could be better explained when first mentioned. When linking, it is unclear on which blocking variables are being used (Section 4.4 & 4.5, page 20). What are you linking with what – full Census to what? Try to be specific and signpost for reader.

1.4 Conclusion:

The panel were content that the method was fit for purpose and suggested some drafting changes to the paper. We would value a more detailed flow chart to be added to the report (see section 2.1 point 1)

NRS post meeting comments

Comments made have been taken into consideration and reflected where appropriate in a subsequent draft.

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: 16th September 2020