Scotland's Census 2022

External Methodology Assurance Panels

Summary Note: Panel 9

Tuesday 23 March 2021

Contents

PSR009: Summary Report of the findings of EMAP Session 9 – Tuesday 23 March 2021

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: -

PMP022: Census - Census Coverage Survey (CCS) Address Linking

PMP023: Household Bias Adjustment

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

Email: censussqa@nrscotland.gov.uk

PMP022: Census - Census Coverage Survey (CCS) Address Linking

Main points of discussion:

The purpose of the paper is to give an overview on the methods for determining whether addresses (which, along with name, date of birth and sex, are the variables used to link both surveys) in CCS and census records are the same or not. This can help inform decisions on whether person records represent the same person. It will also be used to identify which households respond to both the census and CCS, which is used to estimate how many households are missed by the census.

Each CCS address is standardised to correct common address shortenings, remove irrelevant characters and to use consistent address naming conventions where possible then compared to census addresses using a variety of matchkeys with a link between addresses being recorded if there is exact agreement between the matchkeys. The first groups of matchkeys find the most obvious links where a significant portion of the standardised address (such as property, street and postcode, or property information, street and town) are exactly the same. If a link cannot be found using that group of matchkeys then links for the remaining addresses using matchkeys that consist of more limited portions of the address are used, for example only selecting the numbers in the address. Once all of the comparisons have been made, the set of links are collated into one dataset. This dataset of address link information then feeds into the wider estimation process with the address links being clerically reviewed where necessary as part of the Census–CCS person/household linkage.

1.1 Panel members agreed that the methodology was sound. It was expressed that the was clearly written and included helpful appendices and contained useful information on the scale of impacts. An expanded and annotated version of Figure 1 with links to associated papers was suggested.

1.2 The panel requested detail on the consequences of missing data in the CCS or Census. NRS acknowledged that this is dependent on what information is missing though would likely result in a match not being found though locality is more often missing than building names or numbers. There was further discussion on which types of data is most likely to be problematic in terms of matching. It was suggested that flats with different naming conventions are challenging in this regard. One suggested approach was linking the Census to itself.

1.3 There was discussion on how postcodes which had been created since Census day would be handled. NRS suggested this was unlikely as the CCS sample is based on postcodes as at Census day.

1.4 The panel requested more detail on why, as detailed in Section 5.2, both minimal and maximal levels of standardisation are used on addresses. While NRS explained

that applying minimal standardisation of addresses provides additional matches as well as providing additional assurance that matches are correct, additional information was sought on whether this constitutes an additional matching process.

1.5 The panel noted that the quoted figure of 10% unmatched addresses represents a potentially large burden for clerical review however NRS pointed out that this would represent a proportion of the CCS (with a sample of c.50,000) rather than the main Census. Furthermore, additional analysis against electoral data suggested that this proportion may be lower in practice.

1.6 The panel requested more detail around the string algorithm section under Section 5.2, Step 6, perhaps in the form of an appendix.

1.7 There was a discussion around the used of GPS data for linking. NRS noted that though GPS data would be available for field force staff during CCS it would not form part of the dataset so could not be used for linking. It was also noted that there were limitations on using GPS data in urban areas, particularly in blocks of flats.

1.8 The panel asked why the Health Activity Dataset was used as a source for matching. NRS asserted that, as the addresses in the Health Activity Dataset were manually entered this would better reflect the types of responses expected during Census/CCS.

1.9 More clarity was sought in the paper on what was meant by match keys while it was requested that totals be added to tables.

1.10 The panel queried why, as noted in Section 3, the CCS and address frame are different processes. It was explained that the software used to collect addresses is different between the two but that differences between the two would be mitigated through training to field force staff on standardising the format of addresses on collection.

1.11 The panel sought more detail on the impact on clerical review, perhaps by noting the potential extent of clerical review for each matchkey group. NRS sought to look at this in more detail though noted that it was difficult to anticipate in great detail prior to seeing the final data.

1.12 The panel requested more detail on the general point of the specific issues associated with address data and why this methodology was needed, in particular the rationale for grouping was requested. It was also suggested that a comment on the methodology used by other agencies may be worthwhile.

Conclusion:

The panel was broadly happy with the methodology and how it was explained in the paper.

More detail, as far as possible, was requested to be added to the paper around the impact of the matching process on clerical review.

Additional detail on matchkeys and on the issues associated with certain types of address data as well as on why the chosen methodology was used, particularly with regard to grouping.

Additional detail was requested for the flowchart at Figure 1 and for totals to be added to the tables.

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: Katherine Keenan

Date: 23rd March 2021

PMP023: Household Bias Adjustment

Main points of discussion:

This paper explains how the dependence between the census and the Census Coverage Survey is measured. It also details how the estimates can be changed to be more accurate using this information. An alternative estimate of the number of households is needed for this. This estimate is made up of the households that respond to census, as well as a number of households that did not response to the census. The paper details how we decide how many of these households should be included.

Population estimates will be produced for Scotland's Census 2022 using Dual System Estimation (DSE). One of the primary assumptions in DSE is that there is independence between the two systems used in producing estimates – in this case the census and the Census Coverage Survey (CCS). The CCS is designed as an interviewer led survey and held after the census collection period as ended in order to minimise the dependence with the census. However, there will still be some people who do not respond to the Census that are more likely to also not respond to the CCS. This introduces bias to the estimates produced. This paper will detail the statistical method used to correct for this bias, based on the method used in 2011. It will also detail the information needed to implement this correction. This will primarily be focused on between-household dependence, though will also briefly detail within-household dependence.

2.1 The panel agreed that the methodology was sound subject to clarification on the methodological option taken.

2.2 The panel suggested the addition of a flowchart/diagram summarising the process and switching Section 4 and 5 as well as expanding on the implications of the odds ration being less than, equal to or greater than 1. More detail was requested in section 4.2 on how adjustment was done at person-level and of the process in Section 5.7.2.

2.3 The panel requested more description of the process and need for its use and implementation in the introduction. NRS noted that this information was contained in previous papers linked to in footnotes but that more information could be added to set the scene.

2.4 The panel asked if other surveys (e.g. Scottish Household Survey) could be used to illustrate the dependence between individual and household responses. NRS noted that there were difficulties in availability of data and ability to undertake data linkages (especially linking at individual level).

2.5 The panel requested some signposting and clarification on a number of sections including what is meant by in-household and between-household dependence

(Section 3), clarifying the likelihood of the Scenario outlined in Section 6 occurring and adding an explanation of what happens to blank questionnaires in Section 8 and the importance of estimating $n00$.

2.6 The panel requested clarification on which was the preferred option and the reasons for choosing it. NRS noted that this did not appear in the paper at this point as decisions were still to be made in this regard.

2.7 The panel requested more clarification on why the issue of household dependents was not considered a big risk and on the implications of the shift to online returns, including incomplete online returns.

Conclusion:

The panel was content with the paper subject to a decision on which option was being chosen.

The panel requested more detail and signposting in a number of sections including on the reasons for undertaking the methodology, person-level adjustment, in-household and between-household dependence and the importance of estimating $n00$.

More explicit reference to methods used in 2011 was requested.

The panel requested the addition of a flowchart/diagram to summarise the process.

More information was requested on the risks (or otherwise) associated with the issue of household dependents and the shift to online Census completion.

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).	

Chair: Katherine Keenan

Date: 23rd March 2021