Scotland’s Census 2022

External Methodology Assurance Panels

Summary Note PSR003: Panel 3

Wednesday 29 July 2020

Contents

PSR003: Summary Report of the findings of EMAP Session 3 – Wednesday 29 July 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 at this panel session were: -

PMP007: Digital Exclusion Index paper

PMP008: Census Coverage Survey - Communal Establishment Sample Methodology paper

PMP009: Estimation Areas - Geographical grouping for stratification of population estimates paper

Comments and queries welcome to:

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

Email: censussqa@nrscotland.gov.uk

1. PMP007: Digital Exclusion Index paper

Main points of discussion:

The purpose of this paper is to describe the methodology used to create a Digital Exclusion Index.The index partitions small area geographies across Scotland to indicate areas where respondents may be less able or willing to respond to the Census online, in order to direct assistance to where it’s most needed.

Research was carried out into the potential factors underlying digital exclusion, and into what demographic data is available as a proxy for these factors. Data sources were then used to map the geographic distribution of these proxy factors across Scotland, and models were produced to assign digital exclusion scores to each area. The areas were then ranked by the score, and segmented in to five categories for the index. This index was then tested using the 2019 Census Rehearsal data, and improvements were made.

1.1 The panel recognised that this was an impressive and considerable piece of new work, and thought that the approach and conclusions were not unreasonable, though found some parts of the methods unclear. NRS explained that elements in the paper may appear indirect as much of the work was exploring new, uncharted ground, with many approaches reaching a dead end and needing revision.

1.2 The flowchart guiding the reader through the process was appreciated, though it was suggested that numbering each step and using them to structure the paper and signpost to each section.

1.3 There was a discussion around the data sources used, and the timeliness of the OxIS data used as the primary source for creating the model. NRS explained that this research had been carried out in 2018 and explained that there was difficulty in finding alternative sources of information with an appropriate level of geographic granularity. Additional points were raised around internet access through other means than home broadband, if there is data to more accurately reflect access through mobile internet and other means. NRS responded that at that time there didn’t appear to be a good data source for mobile internet, the nearest being based on signal strength which could vary simply by being indoors or outdoors. There are many different ISPs in Scotland, with some specialist rural ones, which makes requesting data difficult. Questions were raised by the panel as to whether the Scottish Household Survey (SHS) data would be more suitable to build the model as the data is more recent. The SHS was tested against the OxIS The panel thought that more explanation around the reasons for using data is needed in the paper.

1.4 There were questions about the variables used in the model, such as disability and education. Similarly, NRS explained that at the time data could not be found that would give this demographic information at a low enough geographic level without it being disclosive.. The panel asked for more clarity around the reason for excluding these variables in the paper. There were also concerns expressed about the level of imputation needed for the income data used in the model.

1.5 There were questions on the choice of PCA (Principal Components Analysis), as well as the specific methods used with PCA. There was some confusion caused by the mix of quantitative and categorical variables, and the transformations to be able to use the data. It was suggested that the correlations may be non-linear between variables. A range of different approaches were suggested

1.6 Some reservations were also raised around the use of the GLM.

1.7 Questions were raised around the approaches in other national statistical agencies. NRS explained the different approaches being taken.

1.8 The panel reflected that the effect of Covid-19 on internet may have an effect on this particular analysis and noted that this was a moving/developing area at the moment. They noted that the research was conducted before 2020.The panel highlighted it as a reason to use more up to date data for the model.

1.9 There was discussion around the content of the annexes, with some material suggested to be included in the main paper, and some background methodology descriptions suggested to go to an annex. Some drafting improvements were also suggested for the paper.

1.10 More context on what interventions would be planned based on the DEI would be appreciated in the paper.

1.11 The panel would like to see a strengthened conclusion section to summarise the strengths and limitations of the work.

Conclusion:

The panel appreciated that the work was in a new area, but had some reservations on the methodology used and the justification within the paper. Changes were recommended to the paper.

Issues were highlighted around the data sources used in the paper, suggesting that either other data sources be found, or more explanation in to why other sources were not suitable (such as SHS which was used to test the model).

The panel were looking for revisions to the choice of PCA and GLM methodologies used. Some alternative approaches were suggested by individual panel members.

Further information regarding the effect of Covid-19 on internet uptake was suggested to be added where possible.

A conclusion with the strengths and limitations of the methods in the paper was requested.

Panel Advice

	Tick('✓')where appropriate
The Panel's advice is that the proposed methodology is fit for purpose (see note below).	✓
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):

The panel are happy to approve methodology with the concerns noted in the meeting and the notes provided in advance. We recommend that the paper includes additional sensitivity analysis to reassure that our suggestions on data and methods in particular would not result in a substantially different index.

NRS Response to EMAP advice:

The creation of a Digital Exclusion Index is being considered further in light of both the panel's feedback and also emerging operational needs.

Chair: Alan Marshall

Date: 25th August 2020

2. PMP008: Census Coverage Survey - Communal Establishment Sample Methodology paper

Main points of discussion:

The purpose of this paper is to detail the methodology for including Communal Establishments (CEs) within the Census Coverage Survey (CCS).

CEs would be sampled within the standard CCS sample areas as in 2011, but with an additional boost sample drawn to avoid the risk of the sample being too small for stratification in the Estimation process. Stratification for the boost sample was explored using geography, CE types or a combination of the two. Creating a sample frame clustered near to the existing CCS areas was also considered, to minimise the operational burden in interviewers travelling to the CEs within the boost sample.

The proposed option for stratification was using Estimation Area and collapsed CE type. While the clustered sample frame was found to reduce travel times, there was concern in the bias it introduced. Therefore the non-clustered approach was proposed, as the change in travel time was not considerable enough to accept the risk of bias.

1.1 The panel agreed that the paper was well written and structured, and the methodology seemed sound with good justification.

1.2 The panel asked for more background on why the sample boost was needed to be added to the paper, and why certain types of establishment were excluded from the CCS.

1.3 There were questions on how Covid-19 would impact on some of these establishments, particularly thinking of the travel, medical and care establishments.

1.4 Suggestions were made around the potential to use historic and seasonal data to inform expected number of residents at Census time.

1.5 Some comments were raised around the quality of the plots, as there is some pixilation. The y-axes of the graph are different between certain graphs, decreasing the ease of comparison.

1.6 There was some confusion around the weighting used to guide the allocation, with respect to size of CE. Suggested that a diagram or flowchart may help explain the process used to calculate allocation to each stratum.

1.7 A question was raised around using response rates for different establishment types from 2011 to inform work. NRS highlighted that in 2011, due to the small sample size that all establishment types were collapsed together, so the data from 2011 cannot be broken down to further granularity.

Conclusion:

The panel agreed that the current methodology appears sound, though the paper would benefit from some clarifications and minor adjustments to the figures.

The paper would benefit from more context on the need for the sample boost on top of the standard drawn sample.

The quality of some of the plots in the paper could be improved, making the axes consistent when they are to be compared, and using higher resolution figures.

Further explanation of the weighting methodology was requested, potentially with the aid of a diagram or flowchart.

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: 25th August 2020

3. PMP009: Estimation Areas - Geographical grouping for stratification of population estimates paper

Main points of discussion:

The purpose of this paper is to describe the methodology used in creating Estimation Areas (EA), groupings of Local Authorities (LAs) for use in Estimation.

Recommendation in previous research was that we did not need to stick to contiguous groups as in 2011. These groupings were previously used to batch data for all data processing (due to computing power limitations at the time), but now would only be used to group data together for Dual System Estimation (DSE) where we think the response profiles would be similar and reduce heterogeneity.

It would be possible to post-stratify to create these groupings, but it would waste the response rate equalisation during the live Census operation, which can be targeted to equalise response within grouped areas. Therefore the areas should be grouped in ways that we predict will be homogeneous in response rate.

A manual approach using the Hard to Count (HtC) scores and index, and a machine learning cluster algorithm, were each used to produce four potential EA groupings. These were evaluated for how much variation there was in HtC scores, the proportion of the population captured in each group, and the likelihood that some variables would require collapsing. Grouping with seven EAs produced by cluster analysis was selected as the proposed option based on analysis.

1.1 The panel found that the paper seemed sound, though noted it was technically more difficult than the previous paper.

1.2 The move to non-contiguous areas seems to be sound. It was commented that there was no comparison of EAs using contiguous areas.

1.3 The panel requested more explanation for some of the terminology used, particularly highlighted were “response rate equalisation” and “downstream processing”. There could also be an earlier definition of planning areas.

1.4 A need for more signposting in the paper, to direct the reader to tables and figures within the text, was raised. The references to each group are not always clear throughout the paper.

1.5 The panel asked for more clarity on checking the numbers of people of non-white ethnicity, as the reason this would require collapsing was not made clear, or why ethnicity was the only variable explored in this way. Collapsing is required to have a sufficient number of people to reduce variance in the estimates when there are small numbers of particular populations. This is most apparent when considering the ethnicity variable, with non-white ethnicities most likely to require collapsing. Other variables used in estimation do not show notable variation between Local Authorities. More detail on this can be added to the paper.

Conclusion:

The panel agreed that the methodology is sound, and suggested some minor improvements that could be made to the paper.

Further explanation of terminology used would be helpful in the paper.

Clearer signposting towards figures and tables would make the paper easier to read. In addition, clearer naming and reference to each EA grouping would avoid confusion between each group.

Post meeting note: since the panel session on 29th July 2020, the estimation areas analysis has been re-run. This is due to the Hard to Count (HtC) index data being updated since the initial analysis was carried out. NRS will add these results into the Estimation Areas methodology paper and will take the paper to Internal Peer Review Group for review prior to publication on our website. Our recommendation will be that the algorithm for producing estimation areas is sound (and was considered fit for purpose by EMAP) so we should use this algorithm for producing estimation areas for 2022. The data will need re-run again prior to 2022 as HtC data is likely to change before live census.

Panel Advice

	Tick(‘✓ ’)where appropriate
The Panel’s advice is to 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: 25th August 2020