

**POPULATION AND MIGRATION STATISTICS (PAMS) COMMITTEE
(SCOTLAND)**

Grouping data for Scotland's Census 2021 coverage adjustment

Scope

1. This paper outlines current proposals for how data from census responses will be grouped during coverage adjustment for the 2021 Census.

Background

2. Every effort will be made during the census to ensure everyone is counted. However, it is inevitable that some people will be missed and they may be missed for a variety of reasons. This results in an under-count.
3. There may also be occasions where people are counted more than once giving an over-count; or they may be counted in the wrong place giving an over-count in one area and an undercount in another. Neither the under-count nor over-count is likely to be uniformly spread across all geographic areas or across sub-groups of the population.
4. As in previous years the 2021 Census will include a coverage adjustment process to measure the population that was missed or over-counted and adjust the population and household estimates accordingly.
5. The coverage adjustment process uses responses to both the census and the census coverage survey¹ to work out the net under-count of people and households for sub-groups of the population. To improve the efficiency of processing the responses are split into a number of groups. Ten groups were used for the 2011 Census. Coverage adjustment is done on a group by group basis. To do this statistical methodologies² are applied using all the data in the group.
6. Analysis of previous census responses has shown that levels of response vary according to a number of demographic characteristics. These include the percentage of the population that are students, levels of migration, level of deprivation, percentage of certain types of accommodation, percentage of private rentals, age group distribution, percentage of urban and rural areas. If these factors vary across different geographies in the group then the expected response levels for those areas will also be different.
7. For example, one area may have a high number of students, which historically have relatively poor response rates, while another area has a very small number of students. When these areas are processed in the same group coverage adjustment is likely to slightly under estimate the number missed in the high student area and

Footnotes

1) The census coverage survey is an independent voluntary doorstep collection survey of around 40,000 households across Scotland. It is undertaken approximately six weeks after census day and asks a subset of the census questions.

2) Dual system estimation is used to estimate the number of missing people and households. The two fundamental underlying assumptions of this methodology are homogeneity of expected response levels and independence between the two surveys used (the census and the census coverage survey).

slightly over estimate the number missed in the very low student area. This is an over simplification as many different factors in each area influence potential response levels but it illustrates the influence grouping can have on the estimates produced. By grouping areas with similar demographics together we can minimise the variation in expected response rates and produce the best estimates for each area.

How data was grouped for the 2011 census

8. Ten groups of data were used during all stages of data processing. Each group contained one or more local authorities (Table 1) with approximately equal estimated populations in each - around 500,000. Where possible each group contained geographically adjacent authorities although one exception was made to allow for practical issues associated with data collection (Shetland Islands were grouped with Aberdeen City and Aberdeenshire). Groups of adjacent authorities had also been used in 2001 (refer to Table 2) although they differed from those used in 2011.
9. The design of the groupings for 2011 focussed mainly on grouping adjacent authorities together as it was felt to be the most appropriate approach to meet early data processing requirements including the handling and scanning of paper forms. It had been noted that regrouping could be undertaken at later stages of processing if different groupings would be better for particular stages of processing such as coverage adjustment. However, no changes were made during live processing for the 2011 Census and the same groupings were used throughout.

Table 1

Local Authority based groups for Census 2011
East Lothian, Scottish Borders, South Lanarkshire
Dumfries and Galloway, East Ayrshire, North Ayrshire, South Ayrshire
City of Edinburgh, Midlothian
North Lanarkshire, West Lothian
Clackmannanshire, Falkirk, Fife
Glasgow City
East Dunbartonshire, East Renfrewshire, Inverclyde, Renfrewshire, West Dunbartonshire
Angus, Dundee City, Perth and Kinross, Stirling
Aberdeen City, Aberdeenshire, Shetland Islands
Argyll and Bute, Highland, Moray, Orkney Islands, Na h-Eileanan Siar

Table 2

Local Authority based groups for Census 2001
Dumfries and Galloway, East Ayrshire, North Ayrshire, Scottish Borders, South Ayrshire
City of Edinburgh, East Lothian, Midlothian, West Lothian
Clackmannanshire, Falkirk, Fife, Stirling
Aberdeen City, Aberdeenshire, Moray
Angus, Dundee City, Highlands, Na h-Eileanan Siar, Orkney Islands, Perth and Kinross, Shetland Islands
North Lanarkshire, South Lanarkshire
East Dunbartonshire, East Renfrewshire, Glasgow City, West Dunbartonshire
Argyll and Bute, Inverclyde, Renfrewshire

10. The groups used in the previous census worked well in terms of allowing data processing to progress in batches and the size of each group was manageable.

Why a different approach is being considered

11. One of the key assumptions of the statistical methodology used to estimate the number of people and households missed during the census is that similar response levels are expected across all sub-groups of people and geographies in the group. For the 2011 Census the variation of the level of response within each group had an impact on the precision of the estimates and the size of the confidence intervals around those estimates³. To ensure we are able to produce the best possible population estimates from the census we are reviewing how the data was grouped to see if alternative groupings might minimise the variation in response rates within the groups.

Proposed approach for grouping data for Scotland's Census 2021

12. Local authority based grouping worked well for the last census so the aim is to improve and refine the approach rather than replace it with something different.
13. Local authorities will be grouped together based on similarity of demographics and expected response rate. The proposed option for doing this is to group local authorities together depending on the percentage of data zones at each Hard to Count⁴ level. Eight groupings have been initially proposed, rather than ten which were used for the previous census, but the number will be reviewed once the 2021 Hard to Count levels have been finalised and the distribution of data zones across the levels in each local authority have been reviewed.
14. The first group contains the three island authorities. The remaining 29 have been split across groups based on the proportions of data zones within the authority at each 2011 Hard to Count level.
15. The first step was to group authorities where more than five per cent of their data zones were classed as hardest to count. As might be expected this identified the four major city authorities. These were split into two separate groups to manage the size of the group. Glasgow City, as the largest, on its own and the other three together.
16. Next the easiest to count authorities were put into a group. A minimum of 60 per cent of data zones at the easiest to count level was selected as the criteria for this group.
17. The remaining 20 authorities were then ranked according to the percentage of data zones in the two easiest to count categories. Determining where to split these was

Footnotes

3) A measure of the precision or uncertainty of estimates is generally provided by estimates of variance and the associated confidence intervals. For the census a bootstrap method using sampling with replacement for 2000 samples is used to estimate the variance of the population estimates.

4) The Hard to Count Index (HtC Index) is an integer classification applied to geographies within Scotland which indicates their expected census household return rates. The classification is derived from the household return rates observed during Scotland's Census 2011 and from the current demographic characteristics of the geographies.

more subjective - for example East Lothian could have been the last authority in group 6 or the first in group 7. The proposed groupings were made taking into account estimated population sizes and to split the estimated population as evenly as possible across groups. The provisional groupings are shown in Table 3.

18. Once 2021 Hard to Count levels are available the same process can be applied and final 2021 groupings determined. Eight groupings have been proposed but the number will be reviewed once the new Hard to Count levels have been generated as the distribution of data zones across the levels in each authority may be better represented by a different number of groups.

Table 3

Proposed local authority based groups for Census 2021 (subject to revision once Census 2021 Hard to Count levels have been finalised)	
Island Authorities	Na h-Eileanan Siar , Orkney Islands, Shetland Islands
Hardest to Count	Glasgow City
	Aberdeen City, City of Edinburgh, Dundee City
↑ Increasing difficulty to count	Fife, Inverclyde, Renfrewshire , Stirling
	Falkirk, North Ayrshire, North Lanarkshire, Perth and Kinross
	East Lothian, Highlands, Scottish Borders, South Lanarkshire, West Dunbartonshire
	Angus, Argyll and Bute, Clackmannanshire, East Ayrshire, East Dunbartonshire, East Renfrewshire, West Lothian
Easiest to count	Aberdeenshire, Dumfries and Galloway, Midlothian, Moray, South Ayrshire

19. Grouping authorities by similarity of distribution of Hard to Count levels should reduce variation in levels of response.
20. Consideration was given to using groups that were not based on local authorities. An option would have been to use groups containing Scotland wide data for a particular demographic. For example all data zones in Scotland where more than ten per cent of the population (as recorded in the 2011 Census) were students, or the least deprived 10 per cent of data zones as identified by the Scottish Index of Multiple Deprivation (SIMD). This approach could also potentially reduce the variation in response levels. However, keeping local authority based groupings makes it easier to take advantage of the existing knowledge base of local authority based demographics and reduces the amount of extra work required to prepare comparator sources. This will minimise time delays, means data quality assurance will be more efficient and local authority level reporting less complex.
21. PAMS committee are asked to note the planned changes and provide any feedback on the proposed change in groupings as outlined in Table 3.

**NRS : Statistical Futures
May 2017**